

Modelos Matemáticos para el Estudio de Medio Ambiente, Salud y Desarrollo Humano

(TOMO 3)



Editores:

Carlos N. Bouza-Herrera

José Félix García Rodríguez

Agustín Santiago Moreno

María del M. Rueda García

*Modelos Matemáticos para el
Estudio del Medio Ambiente,
Salud y Desarrollo Humano.*

Tomo 3

Editores:

Carlos N. Bouza Herrera

José Félix García Rodríguez

Agustín Santiago Moreno

María del Mar Rueda García

Primera edición. Marzo del 2017.

© Universidad de la Habana, Cuba.
Universidad de Granada, España
Universidad Juárez Autónoma de Tabasco, México
Universidad Autónoma de Guerrero, México

ISBN: 978-84-617-8721-0

El contenido de este libro es responsabilidad de los autores.
La presentación y disposición en conjunto de este libro son propiedad de los editores.
Ninguna parte de esta obra puede ser reproducida o transmitida, mediante ningún sistema o método, electrónico o mecánico, sin consentimiento de los autores.

Comité Científico:

Marco Negreiros, Universidad Estatal, Fortaleza, Brazil.

Pedro Monterrey, Universidad del Rosario, Colombia.

Amer Ibrahim Al-Omari / Ph.D. of Statistics Vice Dean of the Academic Research Department of Mathematics, Faculty of Science Al al-Bayt University, P.O. Box 130095, Mafraq 25113, Jordan. Mobile: 00962777906433

Marie Cottrell, Professeur émérite - Université Paris1. Tel et fax(prof): 33 1 44 07 89 22 SAMM, Université Paris 1, 90, rue de Tolbiac-75634, PARIS CEDEX 13-FRANCE. E-mail: cottrell@univ-paris1.fr

Jesús E. Sánchez García, departamento de física aplicada de ICIMAF, la Havana, Cuba.

Minerva Montero Díaz, departamento de física aplicada de ICIMAF, la Havana, Cuba.

Habana, Cuba, Marzo de 2017.

Índice de capítulos

CAPITULO	AUTORES	TÍTULO	PÁGINAS
1	Rilton Primo; Francisco Freires; Marcelo Silva y José Félix García Rodríguez	TRIPARTITE REGULATION OF INVERSE LOGISTICS FOR SOLID WASTE	1 – 9
2	Enrique Vicente Galindo; Belén Vaquero Romero; Purificación Vicente Galindo; Nathalia Tejedor-Flores; Sergio Hernández-González y Purificación Galindo Villardón	ANÁLISIS MULTIVARIANTE DE LA SOSTENIBILIDAD ECONÓMICA, AMBIENTAL Y SOCIAL, EN LAS EMPRESAS EUROPEAS (GRI)	10 – 29
3	Farid García Valentín; Yasser Vázquez Alfonso; Yerandy Carlos Rodríguez Salgado; Jorge Lázaro López y Eddy Pérez Coba.	SISTEMA DE INTELIGENCIA DE NEGOCIO PARA ANALIZAR INFORMACIÓN HIDROLÓGICA.	30 – 36
4	Ignacio Caamal Cauich; Verna Grisel Pat Fernández; David Martínez Luis; José Félix García Rodríguez y Alberto Pérez Fernández.	CRECIMIENTO ECONÓMICO; MIGRACIÓN Y REMESAS EN MÉXICO	37 – 50
5	Carlos Jesús Cruz Bracho; María Amparo León Sánchez; Eugenio Hernández Rodríguez y Víctor Ernesto Pérez León	MODELO DE PREDICCIÓN DE SERIES TEMPORALES PARA LA PRODUCCIÓN DE HUEVOS EN LA EMPRESA COMBINADO AVÍCOLA DE PINAR DEL RÍO.	51 – 63
6	Sahily Estradé Fernández; Vivian Sistachs Vega; Dra. Mirta Álvarez Castilló	ANÁLISIS DE LA SENSIBILIDAD ALÉRGICA OCUPACIONAL EN PANADERÍAS DE LA HABANA	64 – 72
7	Gladys Linares Fleites; María de Lourdes Sandoval Solís; Brenda Catalina Matías Castillo; Hortensia J. Reyes Cervantes; Reyna Almaray y José A. Ticante Roldán	INTERACCIONES MEDICAMENTOSAS: UN ESTUDIO DE CORRELACIONES CANÓNICAS EN PACIENTES PEDIÁTRICOS	73 – 84
8	Agustín Santiago Moreno; José Maclovio Sautto Vallejo; Diego Guzmán-Balderas; Arturo de León Chapa y Jhony Alberto De La Cruz Vargas	ESTIMACIÓN DE RIESGO CARDIOVASCULAR SEGÚN MODELO DE FRAMINGHAM Y MODELO <i>LOGIT</i> EN PACIENTES DE LA JURISDICCIÓN SANITARIA 5 DE GUERRERO.	85 – 108

CAPITULO	AUTORES	TÍTULO	PÁGINAS
9	Luis René Marcial Castillo; Marcela Rivera Martínez; Lourdes Sandoval Solís; Aldo Rafael Matus Angulo; Erika Leonor Basurto Munguía	CIFRADO DE IMÁGENES MÉDICAS USANDO OPERACIONES DE ADN Y MAPAS CAÓTICOS	109 – 116
10	F. D. Martos-Benítez; V. Sistachs Vega; M. A. Díaz Martínez	PREDICCIÓN DE MORTALIDAD EN CUIDADOS INTENSIVOS ONCOLÓGICOS MEDIANTE REGRESIÓN LOGÍSTICA DE COMPONENTES PRINCIPALES CATEGÓRICOS	117 – 128
11	Carlos Bouza; Sira Allende; José M. Sautto y Malay Taylor	UN ANÁLISIS MATEMÁTICO DE EPIDEMIAS TRANSMITIDAS POR EL AEDES AEGYPTI: DENGUE; CHIKUNGUNYA Y ZIKA.	129 – 140
12	Anabel Vega Calcines y Carlos Pérez Risquet	PROCEDIMIENTO PARA LA EVALUACIÓN DEL DESARROLLO SOSTENIBLE TERRITORIAL. APLICACIÓN EN LA PROVINCIA DE VILLA CLARA	141 – 159
13	José Félix García Rodríguez; Carlos M. Lezama León, Ignacio Caamal Cahuich y Rilton G. Bonfin-Primo	EVOLUCIÓN DE LAS CONDICIONES SOCIOECONÓMICAS DE HOGARES BENEFICIADOS CON EL PROGRAMA PENSIÓN PARA ADULTOS MAYORES (PPAM) EN TABASCO; MÉXICO. PERÍODO 2013-2015	160 – 165
14	Pedro A. Monterrey	LA INTEGRACIÓN DE RESULTADOS DE LAS PUBLICACIONES CIENTÍFICAS A PARTIR DE LAS PRUEBAS DE SIGNIFICACIÓN EN EL CASO DE LA PRUEBA PARA LAS MEDIAS DE DOS POBLACIONES	166 – 169
15	Alberto Baena; Beatriz Cobo y María del Mar Arcos	MÉTODOS ALTERNATIVOS A LAS TÉCNICAS DE RESPUESTA ALEATORIZADA PARA ESTUDIAR COMPORTAMIENTOS SENSIBLES EN ENCUESTAS DE SALUD	170 – 178

CAPITULO	AUTORES	TÍTULO	PÁGINAS
16	Jackeline Jiménez; Ma. Paula Iñiguez; Diana Cajamarca; Priscilla Massa Sánchez y Valentín Alejandro Martínez Fernández	ANÁLISIS DE LA HUELLA ECOLÓGICA DEL ECUADOR: UNA COMPARACIÓN CON AMÉRICA LATINA IMPACTO EN LA BIODIVERSIDAD Y LA INCIDENCIA DEL DESARROLLO TURÍSTICO SOSTENIBLE	179 – 187
17	Carmen Viada González; Martha Fors López; Patricia Lorenzo Luaces; Lizet Sánchez Valdés; Mabel Álvarez Cardona; Aliuska Frías Blanco; Lázara García Fernández; Yanela Santiesteban González y Yuliannis Santiesteban González	ESTIMACION DE LA SUPERVIVENCIA EN PACIENTES CON CANCER DE PULMON TRATADOS CON VACUNAS TERAPEUTICAS CUBANAS.	188 – 196
18	Beatriz Cobo, F. López –Torrecillas, D. Molina	APLICACIÓN REAL DE LA TÉCNICA DE HORVITZ DE RESPUESTA ALEATORIZADA AL ACOSO SEXUAL	197 – 202

Afilación

Nombre	Afilación
Agustín Santiago-Moreno	Facultad de Matemáticas, Universidad Autónoma de Guerrero- UAGro. México.
Alberto Baena	Universidad de Granada, España
Alberto Pérez-Fernández	Universidad Autónoma del Carmen, Campeche, México.
Aldo Rafael Matus-Angulo	Benemérita Universidad Autónoma de Puebla, México.
Aliuska Frías-Blanco	Centro de Inmunología Molecular, Habana, Cuba
Anabel Vega-Calzines	Calle 1ra #121 A (altos) e/. D y E, Rpto. Sta Catalina, Santa Clara, Cuba
Arturo de León-Chapa	Universidad Autónoma de Guerrero- UAGro, México.
Beatriz Cobo	Departamento de Estadística e Investigación Operativa, Facultad de Ciencias. Universidad de Granada.
Belén Vaquero-Romero	Universidad de Salamanca España
Brenda Catalina Matías-Castillo	Facultad de Ciencias Físico Matemáticas. Benemérita Universidad Autónoma de Puebla, México.
Carlos Jesús Cruz-Bracho	Universidad de Pinar del Río, Cuba
Carlos M. Lezama-León	Universidad Juárez Autónoma de Tabasco, México. DACEA
Carlos N. Bouza	Universidad de La Habana, La Habana. Cuba
Carlos Pérez-Risquet	Tristá # 510 e/. Amparo y Danielito, CP 50100, Santa Clara, Cuba
Carmen Viada-González	Centro de Inmunología Molecular, Habana, Cuba
D. Molina	Departamento de Estadística e Investigación Operativa, Facultad de Educación, Economía y Tecnología. Universidad de Granada. Campus de Ceuta.
David Martínez-Luis	Universidad Juárez Autónoma de Tabasco - UJAT, México
Diana Cajamarca	Universidad Técnica Particular de Loja. Titulación de Economía
Diego Guzmán-Balderas	Universidad Autónoma de Guerrero- UAGro, México.
Enrique Vicente-Galindo	Universidad de Salamanca España
Erika Leonor Basurto-Munguía	Benemérita Universidad Autónoma de Puebla, México.
Eugenio Hernández-Rodríguez	Empresa Avícola de Pinar del Río, Cuba
F. D. Martos-Benítez	Unidad de Cuidados Intensivos 8B, Hospital Clínico Quirúrgico “Hermanos Ameijeiras”. La Habana, Cuba
F. López –Torrecillas	Departamento de Personalidad, Evaluación y Tratamiento Psicológico, Facultad de Psicología. Universidad de Granada. España
Francisco Freires	Consulting of Centro de Estudios por la Amistad de Latinoamérica Asia y África - CEALA, Brazil
Gladys Linares-Fleites	Instituto de Ciencias. Benemérita Universidad Autónoma de Puebla, México.
Hortensia J. Reyes -Cervantes	Facultad de Ciencias Físico Matemáticas. Benemérita Universidad Autónoma de Puebla, México.
Ignacio Caamal-Cauich	Universidad de Chapingo, México

Jackeline Jiménez	Universidad Técnica Particular de Loja. Titulación de Economía
Jhony Alberto De La Cruz-Vargas	Escuela de Medicina y de la Maestría de Salud Pública de la Universidad Peruana Unión
José A. Ticante-Roldán	Instituto de Ciencias. Benemérita Universidad Autónoma de Puebla, México.
José Félix García-Rodríguez	Universidad Juárez Autónoma de Tabasco - UJAT, México
José Maclovio Sautto Vallejo	Universidad Autónoma de Guerrero- UAGro, México.
Lázara García-Fernández	Centro de Inmunología Molecular, Habana, Cuba
Lizet Sánchez-Valdés	Centro de Inmunología Molecular, Habana, Cuba
Luis René Marcial-Castillo	Benemérita Universidad Autónoma de Puebla, México.
M. A. Díaz-Martínez	CEMAT “Centro de Estudios Matemáticos”, CUJAE “Ciudad Universitaria José Antonio Echeverría”. La Habana, Cuba.
Ma. Paula Iñiguez	Universidad Técnica Particular de Loja. Titulación de Economía
Mabel Alvarez-Cardona	Centro de Inmunología Molecular, Habana, Cuba
Malay Taylor	Taylor Medical Institute, Kolkata. India.
Marcela Rivera-Martínez	Benemérita Universidad Autónoma de Puebla, México.
Marcelo Silva	Consulting of Centro de Estudios por la Amistad de Latinoamérica Asia y África - CEALA, Brazil
María Amparo León-Sánchez	Universidad de Pinar del Río, Cuba
María de Lourdes Sandoval-Solís	Facultad de Ciencias de la Computación. Benemérita Universidad Autónoma de Puebla, México.
María del Mar Arcos	Hospital de La Paz
Martha Fors-López	Escuela de Medicina, Universidad de las Américas, Quito, Ecuador
Mirta Álvarez-Castilló	Hospital “Calixto García”, Cuba
Nathalia Tejedor-Flores	Universidad de Salamanca España
Patricia Lorenzo-Luaces	Centro de Inmunología Molecular, Habana, Cuba
Pedro A. Monterrey	Departamento de Matemáticas, Universidad del Rosario, Bogotá, Colombia
Priscilla Massa-Sánchez	Universidad Técnica Particular de Loja. Departamento de Economía.
Purificación Galindo-Villardón	Universidad de Salamanca España. ESPOL. Guayaquil. Ecuador
Purificación Vicente-Galindo	Universidad de Salamanca España
Reyna Almaray	Facultad de Ciencias Químicas. Benemérita Universidad Autónoma de Puebla, México.
Rilton G. Bonfin-Primo	Centro de Estudios por la Amistad de Latinoamérica, Asia y Brasil (CEALA), Salvador Bahía, Brasil.
Sahily Estradé-Fernández	Universidad de La Habana, Cuba.
Sergio Hernández-González	Universidad Veracruzana. Xalapa, México
Sira Allende	Universidad de La Habana, La Habana. Cuba

Valentín-Alejandro Martínez-Fernández	Universidad de A Coruña. Departamento de Análisis Económico y Administración de Empresas.
Verna Gricel Pat-Fernández	Universidad de Chapingo, México
Víctor Ernesto Pérez-León	Universidad de Pinar del Río, Cuba
Vivian Sistachs-Vega	Facultad de Matemática y Computación, Universidad de La Habana. La Habana, Cuba.
Yanela Santiesteban-González	Centro de Inmunología Molecular, Habana, Cuba
Yuliannis Santiesteban-González	Centro de Inmunología Molecular, Habana, Cuba

PRÓLOGO

En el mundo, la crisis económica asociada al fenómeno de la globalización ha impactado negativamente en las condiciones económicas y de bienestar social de los países, sean pobres o ricos. Particularmente, en América Latina, la crisis ha acentuado los problemas estructurales asociados al subdesarrollo, como son desigualdad, pobreza, migración, inseguridad y falta de acceso a la salud y la educación.

El tema de salud humana es relevante, pues además de constituir un valor intrínseco para la calidad de vida, es componente fundamental de la productividad de las personas y el capital humano de los países, razón por la cual tiene también un valor instrumental para el crecimiento y desarrollo económico de los países.

Por otro lado, la baja productividad de las empresas, el estancamiento económico y el agotamiento del modelo exportador de materias primas, característico de nuestros países, han generado desempleo, caída del poder adquisitivo e inflación, sin dejar de lado el tema del impacto ambiental, toda vez que la sociedad observa con preocupación creciente la destrucción del medio ambiente derivada de la explotación desenfrenada de los recursos naturales.

En este contexto, son tres los grandes escenarios que deben ser estudiados en nuestros países: la inserción de las economías locales al contexto de la globalización en términos de ventajas competitivas; el estancamiento del desarrollo económico y social y la consecuente caída en la calidad de vida de la población, y el acceso de la población a los satisfactores básicos para la vida, como son salud, educación, vivienda y alimentos suficientes y de calidad.

En las líneas de investigación posibles ante dichos escenarios, debe destacarse el papel activo de los principales involucrados. En primer lugar, el Estado nacional como principal proveedor de bienes y servicios primarios como son salud, educación y seguridad social; en segundo lugar, la empresa y su importante función económica y social, y en tercer lugar la universidad como generadora de conocimiento científico, formadora de capital humano y conciencia crítica de la sociedad. Estado, empresa y universidad están obligados a trabajar de manera conjunta y organizada a fin de definir e instrumentar políticas económicas y públicas orientadas al cambio y el bienestar social.

Evidentemente, el papel activo del Estado como promotor del desarrollo económico y social constituye una de las razones que explican el alto nivel de bienestar y prosperidad alcanzado por los países ricos. Ello ha significado largos períodos de esfuerzos y sacrificios, esfuerzos que han descansado básicamente en el ahorro interno, en la inversión en capital y tecnología, y sobre todo en la inversión pública realizada para la formación de capital humano y físico.

Dicho esfuerzo ha estado acompañado por la presencia de empresas con enfoque de responsabilidad social y comportamiento ético empresarial, así como de la universidad como espacio propicio para la generación y transferencia del conocimiento. En contraste, los países pobres y en vías de desarrollo no cuentan con la suficiente base económica y voluntad política que les permita invertir los recursos necesarios para la superación de los problemas estructurales que los aquejan.

Tal como establece Piketty en su renombrada obra *El capital en el siglo XXI*, el principal mecanismo de convergencia hacia la reducción de las desigualdades en una sociedad es la divulgación de los conocimientos derivados de la investigación científica, considerada ésta como un bien público. Este libro, integrado por diecisiete capítulos producto de investigaciones cuantitativas interdisciplinarias tiene ese propósito.

Los investigadores que participan en la presente obra, de diversas formaciones académicas y diversos países, conjugan sus mejores esfuerzos para estudiar y proponer soluciones desde el ámbito de la investigación de operaciones a diversos problemas relacionados con el bienestar, la economía y el medio ambiente de nuestros países. Los lectores tendrán la última palabra.

José F. García Rodríguez

TRIPARTITE REGULATION OF INVERSE LOGISTICS FOR SOLID WASTE¹

Rilton Primo*, Francisco Freires*, Marcelo Silva*, & José F. Rodríguez**

*Consulting of Centro de Estudios por la Amistad de Latinoamérica Asia y África - CEALA, Brazil.

**Universidad Juárez Autónoma de Tabasco - UJAT, México.

ABSTRACT

The management of reverse logistics costs of solid waste, for alternative uses, is challenging. In these markets, the presence of actors with asymmetrical powers, who generate distributive and environmental inefficiencies, is common, increasing the risk of soil and water contamination by urban and rural waste. That inspires regulations/links between companies, government and society, through bipartite or tripartite governance, these latter recommended here. To manage costs we propose the model Activity Based Costing/Management – Integralized (ABC/M-I), transversal to the Direct Product Profitability (DPP), Total Cost of Ownership (TCO) and Customer Profitability Analysis (CPA). Starting from case studies in both hemispheres, the model is useful to private and mixed networks with different levels of technological incorporation.

KEYWORDS: Inverse Logistics. Solid Waste. Costs Management.

RESUMEN

El manejo de los costos de la logística inversa de los desechos sólidos, es un reto. En estos mercados, la presencia de actores con poder asimétrico, que generan ineficiencias distribucionales y medio ambientales, incrementan los riesgos de contaminación de suelos y agua por los desechos urbanos y rurales. Esto inspira conexiones/regulaciones entre compañías, gobierno y sociedad, a través de gobernabilidad bipartita o tripartita, estos últimos recomendados aquí. Para manejar costos proponemos el modelo “Activity Based Costing/Management – Integralized” (ABC/M-I), transversal a el “Direct Product Profitability” (DPP), “Total Cost of Ownership” (TCO) y “Customer Profitability Analysis” (CPA). Comenzando con el estudio de casos en ambos hemisferios, el modelo es útil para redes privadas y mixtas con diferentes niveles de incorporación tecnológica.

1. INTRODUCTION

Within this 5th World P & OM it might not be casual that three authors from Salvador, Bahia’s state capital, in Northeastern Brazil, R. Primo, F. Freires and M. Silva, and a Mexican from Xalapa, Veracruz, J. Rodríguez, sign this study on the cost management of reverse logistics of solid waste. This short preamble exposes its causality.

This happens not only because the Mexican has kept active networks of scientific cooperation with Salvador and Havana, neither because he has been dedicated specifically to this topic, nor because some professors at Faculty of Mathematics and Computation of Universidad de La Habana (UH) have been working on the increased cost of managing the reverse logistics of solid waste in Havana, with impacts on the environment of what they consider to be of “vitaly important the development of mechanisms to lower costs” (Amos, Alonso, 2014). Looking closer, would it be casual that this same study has been published in the second volume of the book *Experiencias en la Modelación de la Toma de Decisiones en la Salud Humana, Medio Ambiente y Desarrollo Humano*, work that was coordinated precisely by this Mexican?

Improving the question, might it not be contingency that, on the first volume of such publication Rodríguez signs with Primo a study about the distortions of the costs of the chain of reverse logistics of solid waste in Salvador and metropolitan regions of Brazil (Primo, Rodríguez, 2014), which markets are oligopsonic (few buyers and many suppliers), generating environmental damages and social iniquity?

On one hand, J. Amos and S. Alonso note that La Habana generates around 17.000m³ of solid waste daily, and that, because the country is still economically and technologically blocked, it mixes animal and motorized traction in this reverse logistics chain. Considering only the fuel costs and the charge capacity of the trucks, one hundred daily trips and 700 liters are needed, from the collection points until the established spillways, which is a cost directly connected to the distances. Once the truck, fuel and waste offer is kept unchanged, it fits to shorten the routes as an ecologically positive economy, since 800g of CO³/liter of fuel are produced. The authors report that, in recent years, other countries have felt forced to evaluate their solid waste management, examine its cost-effectiveness and impact on the environment in terms of routing designs and collection. They concluded that the *routing design* for the waste

¹ This study was approved by the Scientific Committee of 5th World Conference Production and Operations Management (5th World P & OM) comprised by EurOMA (European Operations Management Association), POMS (Production and Operations Management Society) and JOMSA (Japanese Operations Management and Strategy Association).

collection problem can be developed by using Graph Theory, in which all arcs or edges of an oriented, not-oriented or mixed network are traversed; subject to vehicle capacity restrictions, and the goal is to minimize the total cost of the routes traversed by the waste collection vehicles.

On the other hand, Primo and Rodríguez point out, for Salvador, precisely the strategic decisions of location of the collection points, according to local traffic engineering that minimized the routes that united the first links of the chain to the last ones. One could say that they are “brother-studies”, but the authors do not even know each other.

This riddle was partly explained on the morning of May 14th, 2015, when Salvador’s City Council held a special session that discussed the links between this city and Havana, two metropolises that, since 1993, by a law enacted by the city of Salvador, had been recognized as sister cities. Nevertheless, would the African-religious, cultural and port links, often indicated between Salvador and Havana, be enough to end the mystery from which emerged this pair of convergent, but so far apart studies?

The key to the puzzle would be in the annals of that Special Meeting. These words were heard from the Consul of Cuba in Salvador, Laura Pujol, with implicit contents: Cuba maintains "a socialist model of development which now enjoys the support of Iberian-American countries" (Salvador, 2015). Will our puzzle have its final nexus across the ocean? The reader may evaluate if the fitting of the piece needs to be any tighter.

In the context of academic approaches, there had been signed on March 14, 2014, in Havana, the agreement of the publication of volumes with the two mentioned studies on reverse logistics of solid waste. This decision had been taken at the first meeting of the *Red Iberoamericana de Estudios Cuantitativos Aplicados* – RIDECA until then composed by the Universities of Granada and Autónoma de Guerrero (División Mathematics of Academic's Acapulco), Havana (Faculty of Mathematics) and Juárez Autónoma de Tabasco (División Academic of Administrative and Economic Sciences), then expanded precisely by organizations in Brazil, Spain, Cuba, Ecuador, Mexico and Colombia.

Rideca comes to the fifth P & OM as a planned network, through basins of fractal attractors². Returning to the Iberian Peninsula, more accurately to the University of Porto in Portugal, an independent line of studies on reverse logistics of solid waste was covered up to these lines by one of the Brazilians who signed this article, F. Freires. This happened a decade and a half after his approach to the eminent Cuban production engineer Carlos Taboada Rodríguez, whose work has an emphasis on logistics. As a relevant basis for this study, Freires (2000, 2006, 2007, 2008) describes and analyzes the influence of power and the trust among players over performance (effectiveness and efficiency) of Reverse Logistics System (RLS), indicates environmental externalities and legal restrictions able to redefine locally adapted ways of managing, generating results from which we have developed the overall integrated management model presented here, at last.

Such engineering studies were connected in this article with the reverse logistics policy economics of solid waste. This line began to be covered in 2009, with the advisory services of R. Primo in the Executive Coordination of *Superintendência de Economia Solidária* of Government of the State of Bahia, leading to academic results in book form (Primo, 2012a), articles (Primo, 2012b, 2014, 2015; Primo, Rodríguez, 2016a) and book chapters (Primo, Rodríguez, 2015, 2016b), by stimuli or articulations of the doctoral teams in Political Economy of UH, RIDECA, UJAT and CEALA.

We bring this debate to the fifth World P & OM with echoes of ideas recently heard between 16th and 19th, May, 2016, in *Palacio de Convenciones de La Habana*, at the International Congress on Economic Management and Development. In this moment, Primo and Rodríguez (2016b) presented a study on the macroeconomic context in which the present subject is inserted, the global crisis. The 1st International Meeting on Managerial Direction and Management was part of this Congress, in which Scientific Program the reverse logistics and the sustainable approach for logistics processes and supply chain, urban logistics, information systems and logistics control were specifically discussed.

Beside this introduction, this paper is divided into three sections, dedicated to three phases of the cycle here examined: the first on the integrated management of logistics costs from the beginning of the chain; the second on the administrative arrangements of these costs over the chain and the third on the final cost/price according to the destination.

² The first four editions of the P&OM were held in Seville (2000), Cancun (2004), Tokyo (2008) and Amsterdam (2012). Havana was chosen to host the 5th P & OM, which took place from 6 to 10 September 2016, attracting speakers from Indiana University (USA), University of Minnesota (USA), University of North Carolina (USA), University College Dublin (Ireland), Università di Bergamo (Italy) University of Warwick (UK), University of Texas (USA), Gakushuin University (Japan), Technische Universität (Germany), McGill University (Canada), Universitaet Klagenfurt (Austria), INSEAD (France), Vlerick Business School (Belgium), University of Cambridge (UK) and others.

2. SECOND PHASE: THE RUB.

It is imperative that we leave the technical issue of the third stage, i.e. on alternative uses of solid waste for energy generation, recycling and adequate final destination, to the next section. Let us head straight to the technical issue of the second phase, administrative, that is the rub. Distinct local regulatory environments regulate the degree of selectivity in the collection of waste, providing (or not) equitable gains, industrial uses and ecological final disposal. Unfortunately, in many cities the irrational and even illegal disposal of waste that may be useful or not, or is harmful, the wastage of an important portion with energy value or industrial recycling potential, a socially unjust distribution of the portion that is recovered and an unintegrated or arbitrary management of the costs of reverse logistics of solid waste are common.

As if the problems were not sufficiently challenging, the chain of reverse logistics of solid waste often takes the form of oligopsonic markets and distorts the costs/prices. Such distortions reverberate throughout the chain causing social and environmental externalities (in here, increasing the risk of soil and water contamination by urban and rural waste). Of course. If the reusable solid waste, industrially recyclable or power generators, are collected by a segment or more and the benefits are earned by the other(s) segment(s), the result tends to be economically and environmentally ineffective and socially spurious, creating the urgency of regulation or deliberative equivalence between the actors.

In environments where the government prevails or rules, deadlocks like these are solved more easily with the formation of bipartite (municipal and state) or tripartite (municipal, state and federal) disciplinary workplaces, which are able to plan and coordinate the flows, joint actors, supervise the efficiency, effectiveness and split costs. This type of solution is often considered vertical, because it promotes the reintegration of hierarchized powers. Also outside of the socialist economies this type of solution, among public managers, of operational deadlocks, has been observed, e.g., in the environmental and health sectors in Brazil, with the creation of the Bipartite and Tripartite Committees (CIB and CIT) with articulatory and agreement functions. They are composed by equal numbers of representatives from each public sphere. It is worth stressing that in health CITs the decisions are taken by consensus and not by vote, requalifying this solution from *vertical* or *hierarchical* to *commissioned-argumentative*.

On the other hand, where the private sector is predominant, via industrial oligopsonies and intermediaries, as the recycling of solid waste market in Brazil, the deadlocks are reparable in a less stiff form, also through the creation of some bipartite or tripartite workplace, respectively, integrating public authorities and companies or these actors and organized civil society. It is qualified as a horizontal solution if and only if there is parity that means, if the representatives have equivalent voice and management rights and equal decision-making power. In these atmospheres the tripartite-parity solutions are less worse, despite the old possibility of government *capture* by private interests (Stigler, 1971; Krueger, 1974a, 1974b;), in which case, if there is no consensus, although there is representative parity, social and environmental interests can still be subordinated into having only 1/3 of the votes. Possible arrangements shielded against the consequences of the capture are less feasible. This seems to be a matter that is not only technical, but regarding a breach of trust in the relationship. In this case, a joint management based on the trust between the players is only sustainable up to the point where all players are satisfied. In the other situation, in the management of power of a single player, the maintenance of the relationships is dependent on the motivation of this player in giving continuity (Freires, Guedes, 2008). An effective alternative is the establishment of a local regulator with public support to encourage and guide their action towards the social, environmental and strategic interest.

Here is an abstract scheme of escape routes to the logistics, abstracted from local labyrinths. This article was derived from a series of case studies, but it does not dismiss reviews. Only by immersion in the institutional architecture can an analyst unveil his wiles. A local example of dialogue between interests is the National Tripartite Commission - CTPP (employers, workers and government), restructured by the government of Brazil in 2008 to take into account the manifestation of society on issues related to safety and health at work. Before that, such issues were decided without major concerns over what society thought about them.

Inversely, a local and evil example of bipartite cooperation, with clues of capture, has been provided by the Ministry of Environment of Brazil, responsible for the Executive Secretariat of the Guidance Committee for Implementation of Reverse Logistics Systems (*Comitê Orientador para Implementação de Sistemas de Logística Reversa* - Cori). Cori claims to have focused efforts on the development of sectorial agreements with entrepreneurs in order to implement the shared responsibility for the life cycle of products. They argue that they proceed according to Law nº 12.305/2010 establishing the National Policy on Solid Waste – PNRS and to Decree Nº 7.404/2010 which regulates it, but by both hold the collectivity responsible and, in a chained and individualized way, also hold the Third Sector and the consumer responsible – not only the State and business sectors – for the product life-cycle management and the effectiveness of actions to ensure compliance with the PNRS and the guidelines and criteria established by this

Law and this Decree. According to Art. 11 of the Law of PNRS, the selective collection system should prioritize the participation of cooperatives or other forms of collectors' association of reusable and recyclable material, consisting of individuals of low income. This legal item makes practical sense, because these agents collect about 90% of recycled material. (Freitas, Fonseca, 2012). But up to 75% of this supply chain gains are appropriated by oligopsony (Grimberg, Tuszal, Goldfarb, 2004).

Cori published ten executive deliberations, some dealing specifically with the economic viability of reverse logistics of solid waste, citing the preservation of competition among entrepreneurs: it fell silent on oligopsony and negative externalities (increasing the risk of soil and water contamination by urban and rural waste). It was and still is the historic moment for the state to mediate the oligopsony's contracts with cooperatives, oversee prices, based on undistorted and efficient cost, but it has remained off the scene and has advocated, in official acts, that the reference prices of these contracts will be those in effect in the market, prices at times crippling, inhibiting the lower links in the reverse-logistics chain, abuses denounced by studies (Martins, Peres, 2006; Magera, 2008; Bosi, 2008; Carmo, 2009, 2011; Primo, 2012, 2015; Primo, Rodríguez, 2015, 2016).

To Freires and Guedes (2008), in general, the dilemma of who must manage a logistics system is based on authority and control of the system. This can be of different natures (technological, financial, domain of the market, etc.). The problem of power, which although being a mechanism of control, will be contrasted with what generally is conceived as a mechanism of competitive control between players, trust. These updated perceptions dates back to the origins of regulatory theory (Stigler, 1971, Posner, 1974, 1975).

When Cori excludes the third sector and consumers of the agreements, it fails to pursue the goals of PNRS, in addition to disregarding the resolutions of the 4th National Conference on the Environment, involving 200,000 managers and whose debate "was based on the concept of Shared Responsibility [SR], where each sector of society has responsibilities and roles to solve or mitigate problems related to solid waste" (Brazil, 2015). Such SR is a 'Priority Action' belonging to the 'Axis 2' of this Conference. Not by chance: the bipartite sectorial agreements rehearse a legal, ethical, logistical and ecological regress. Case studies concluded that there is a relation between the consolidation of relationships between players and volumes exchanged in the system (Freires, Guedes, 2008). Thus far only five state-company agreements were signed (with firms in the plastic packaging sector and in general, fluorescent lamps, electronics and medicines) faced with national losses estimated at US\$ 2.2 billion/year with the difference between the volume of recyclable and recycled Municipal Solid Waste (IPEA, 2010), including energy diseconomies necessary for the extraction and processing of raw materials. But we need to add to these accounting losses the economic or opportunity cost of non-use of waste for energy generation. For only two cities of the country (Nova Iguaçu and Rio de Janeiro, whose gravimetric characterizations are close to the national average) such wastage/year is equivalent to 106,835.23 MWh or to 594,000 homes' electricity bills.

Let us focus on this sum the critical argument for horizontal tripartite regulation of logistics chains are highly concentrated, feeders of iniquities and ecological damages.

A basic tool to measure the degree of concentration is the Herfindahl-Hirschman Index:

$$H = \sum_{i=1}^n S_i^2 \tag{1}$$

where S_i stands for the market share (the percentage of the market/industry of each of the i -th firms), n for number of firms, $0 < \alpha \neq 1$. H ranges from $1/N$ to 1, generating *Table 1*:

Table 1 - Spread Index

<i>Herfindahl-Hirschman Index</i>	<i>Market Concentration</i>
H < 0,01	Highly Competitive
H < 0,15	Non Concentrate
0,15 < H < 0,25	Moderate Concentration
H > 0,25	High Concentration

Many regions are between the latter two cases, but do not yet know.

3. WASTE FOR ENERGY USE.

Let us be brief in reviewing the viability of reverse logistics of waste for energy use, which depends on local access to technology and on a certain amount of investment. These characteristics make the abstract recommendations only starting points to the projects. In summary, the most viable ways are



incineration and anaerobic digestion of waste. Viability arises of the portions of Cost Benefit Index - ICB (Tolmasquin, 2003):

$$I = \Pi + K + \Lambda + M \tag{2}$$

where I is the ICB and Π the annual cost of investment at a power plant, see equation 3, where α is the amount of investment for the plant, including the cost of any interest on; in it, β is the capital recovery factor for the economically useful life of the plant, given by equation 4, where i is the discount rate, v the useful life (in years); γ is the guaranteed energy of the plant, measured in MW/year and δ the average number of hours of the year.

$$\Pi = \frac{\alpha\beta}{\gamma\delta} \tag{3}$$

$$\beta = \frac{i(1+i)^v}{(1+i)^v - 1} \tag{4}$$

K is the annual cost of operation and maintenance at the plant, measured in \$/MWh, see equation 5; Λ is the annual cost of investments in transmission, see equation 6; M is the annual fuel cost of the plant in \$/MWh, see equation 7 (Henriques, 2004):

$$K = \frac{\xi\theta}{\gamma\delta} \tag{5}$$

$$\Lambda = \frac{\sigma\omega}{\gamma\delta} \tag{6}$$

$$M = \psi\eta \tag{7}$$

where ξ is the annual cost of operating and maintaining the plant, measured in MW\$/year and θ the power installed at the plant in MW; σ is the total transmission investment, including interest during construction and ω is the capital recovery factor for the economic life of the transmission, ψ is the unit cost of fuel gaugeable in \$/tonne (ton) or \$/m³ and η is the average specific consumption of the plant in ton/MWh or m³/MWh.

Posting the local data on this methodology, the best proposals will be prioritized. Having done this, you may want to define cities' portfolios for incineration, biogas or recycling. For each use, in every city, there are useful methods for viability assessing. As an example, we can adopt the hypothesis of the feasibility analysis of the solid waste incineration for electricity generation coming from a supplier x that does not segregate waste. The viability depends on the efficiency of heat conversion (20% on average) and the calorific value of the material. This last unknown is calculable by equation 8, statistically elaborated by N. Themelis (2003) after case studies:

$$\Omega = \frac{18.500\phi - 2.636\Delta - 628\varepsilon - 544\mu}{4.185} \tag{8}$$

where Ω is called Inferior Calorific Power, usually expressed in kcal/kg; ϕ is the proportion of fuel (putrescible, paper, wood, leaves etc.) in 1 kg of MSW, discounting this calculation the weight of water, whose ratio is measured by the variable Δ (being 60% of the average case) ε is the ratio of glass and μ of metals, resulting in *Table 2*.

Table 2 - Viability Study

Inferior Calorific Power	Diagnostic
$\Omega < 1.675$ kcal/kg	Incineration is not technically feasible (in addition to technical difficulties, requires adding fuel).
1.675 kcal/kg $< \Omega < 2000$ kcal/kg	The viability of incineration still depends on some kind of pre-treatment to raise the calorific value.
$\Omega > 2000$ kcal/kg	Gross burning ("burning mass") is technically feasible.

An unserviceable for incineration part can still be recycled to generate biogas or methane etc. In the US, Massachusetts and Connecticut use their solid waste more to generate energy than recycling; in Minnesota and Maine the opposite occurs. Brazilian studies show that the strategic decision to recycle or generate energy depends on factors related to the supply chain of waste, since the origin of the collection, degree of selectivity, up to pre-treatment for industries or plants (Soares, 2011). In US it was noted that there is a strong correlation between population density and energetic use of waste: the 15 states with the highest density (163 people/km² on average) generate 85% of energy obtained from solid waste and the

ten least populated only 4%; the national average of waste use for generating energy at the beginning of the century, was almost 10% (Themelis, Kaufman, 2004); today, there are cities that tend to 50% by reverse logistics.

Recent research from authors such as M. Silva (Silva et al, 2014a, 2014b), the third Brazilian author of this article, shows that the world experiences a "significant increase in investment in renewable energy," highlighting among them the "biogas (sewage, garbage and manure)". Lastly, it provides new tools for analysis/diagnosis of the competitive character of these markets which include "institutions, mills, cooperatives," methodology that managers "can apply, enabling market players involved can be inserted successfully in the new competitive environment" (Silva, Teixeira, Torres, 2015).

4. FIRST PHASE OF THE LOGISTICS

Let's discuss now some tools of management of the first phase of the logistics. Suppose a chain with any degree of complexity and technological incorporation. Let's keep implied, on account of being an exception, the assumption of logistic chains purely public or state's. Whereas the efficiency and success of the supplier depends on the costs that will be incurred as the waste moves through the logistics system (Christopher, 1997), after a brief analysis we realized that the focus of their accounting shall be on the net contribution by the residue after sales, once added the surcharges and subtracted all the expenses that can be rationally allocated or assigned to waste individually. Such approach is today called Direct Product Profitability (DPP) (Bornia, Freires, 2003).

This is not the only possible balance: it does not cover all cost aspects and the shadow costs. In a context where social and environmental values are central, non-financial variables are even clearer, but they are significant even focusing strictly on the technical bias. Furthermore, the integration of these aspects tend to be increasingly imposing, not an option. For example, the Resolution N. 3 of Cori, August, 24th, 2011, which establishes the criteria and the content of the studies for the approval of the Technical and Economic Feasibility of Reverse Logistics Systems Implementation, brings in its Art. 5: the assessment of the costs of implementation of the post-consumer reverse logistics systems must cover three aspects: I - the direct and indirect economic impacts on other chains; II - the correlation between the market price of the virgin and recycled raw material; and III - the environmental impacts of the implementation and operation of reverse logistics systems, including the increase of greenhouse gas emissions - GHG.

Certainly, the mere DPP fits the case of medium cooperatives, when feasible. It still applies to cases of concession of public services to private collection companies. But it occurs frequently that at the base of the chain small actors and individual actors survive.

A more accurate balance of a chain which base is heterogeneous and / or atomized is known as Customer Profitability Analysis (CPA), tool for which many other costs should be considered before defining the efficiency / viability of certain chain links. In a network with thousands of suppliers of waste it would not be possible to individually examine their efficiency, but it would be possible to select a representative sample in order to obtain an overview of the costs associated with different types of agents, flow channels and specific segments. In general, from CPA the most vulnerable links, ready to break, are revealed. In general, they are deficient in the planning stages of transport, loading, handling, unloading. They overload the chain with above average costs to maintain inventory, to maintain physical warehouse space, materials handling, to manage and realize the internal transport of materials.

A third balance of reverse logistics of waste is the Total Cost of Ownership (TCO). This approach is applied to determine the efficiency of the previous links when the subsequent ones expect from them commodity with specific characteristics, noting potential inadequacy. These should be included in the analysis of the total cost of acquiring a residue or service. The TCO considers costs with research and continuous qualification of the supplier, shipping and receiving, inspection, rejection and expenses associated with correction and inadequate remittances (Ellram, 1995). TCO uses the Supplier Performance Index (IDF), by this formula:

$$IDF = (\text{Costs of inadequacy} + \text{Purchase Price}) / \text{Purchase Price} \tag{9}$$

The following exercise illustrates the calculation of the IDF of a supplier in a certain time:

<i>Spent with acquisition of the solid residues:</i>	\$ 250.000
<i>Costs of inadequacies:</i>	
Return of the remittances by unacceptable pattern: (2 occurrences x \$300).....	\$ 600
Correction of the Imperfect Remittances but acceptable: (5 occurrences x \$350).....	\$ 1.750
Delays in waste delivery: (3 occurrences x \$500)	\$ 1.500

Total Costs of inadequacies:..... \$ 3.850
 $IDF = (\$3.850 + \$ 250.000) / \$250.000 = \$ 1,015$

The evaluation/selection of providers A and B may be performed with this calculation:

	<i>Provider A</i>	<i>Provider B</i>
Unit Purchase Price (UPP)	\$ 100	\$ 105
x IDF	1,1	1,0
Total Cost Unit	\$ 110	\$ 105

Analyzing only the UPP, the conclusion would be that the supplier A is more attractive than B. However, from the perspective of TCO, the supplier B becomes better than A.

Two management tools transverse to all cited above are the Activity-Based Costing (ABC) that, in this context, catch all the indirect costs that are not incorporated into the waste but in operational and management activities of these goods and services. This model was developed by professors from Harvard Business School (Kaplan, Cooper, 1998) and modified by Computer Aided Manufacturing International (CAM-I) (Börjesson, 1997), whose vision is horizontal, focused on those indirect costs incurred by processes chained together of each link in the chain, called Activity Based Management (ABM). ABC and ABM are similar and can be confused, but have distinct contributions.

On one hand, the DPP, CPA, TCO, and on the other, ABC and ABM are tools often not present in a combined way, not allowing the full exploitation of the managing potential of existing costs in the supply chains in which they apply. In the solid waste, the integration of these tools represents a major impact on the visibility of the incurred shadow costs. The integrated model is described in detail in Freires (2000), now summarized and transformed in a transversal administrative model.

Central axis capable of integrating DPP, CPA e a TCO, the extracted data through this system (we can call it ABC/M-I or Activity Based Costing / Management - Integralized) are more reliable than those provided by traditional costing systems. Observing the upper flow of the chain (recycling industries, power plants, cooperatives for collection and pre-treatment and service companies), one can establish non-financial procedures to assist the recipients in evaluating the performance of their suppliers, transporters and supply channels, eliminating activities as reception, requisition and control orders, quality control of received lots, among others. The reduction in direct and indirect costs will also be reflected in the transverse relationship of recipients with previous links.

5. FINAL CONSIDERATIONS.

We must conclude warning that the incentives for the adoption and implementation of these tools are not always clear and useful. Inefficient accounting and economic situations, cost distortions, social inequities and environmental externalities (increasing the risk of soil and water contamination by urban and rural waste) can survive any logic and effort if institutional arrangements keep up irresponsible positions, opportunistic behavior or captured rationality despite the possibilities and demands of increased use of the solid residues for recycling and energy generation, environmentally appropriate disposal, when it is impossible to reuse and reduce them.

REFERENCES

- [1] Amos, J. and Alonso, S. (2015): El Diseño del Recorrido para la Recogida de Residuos Sólidos. Caso de Estudio: Municipio Plaza de la Revolución. In. Herrera, C. et al. (Ed.) Experiencias en la Modelación de la Toma de Decisiones en la Salud Humana, Medio Ambiente y Desarrollo Humano, Tomo 2, Universidad de Granada, Granada.
- [2] Bornia, A.; Freires, F. (2003): Uma aplicação do custeio baseado em atividades à gestão dos custos da cadeia de suprimentos In: VII Congreso del Instituto Internacional de Costos, Punta del Este.
- [3] Bosi, A. (2008): A Organização Capitalista do Trabalho ‘Informal’: O caso dos catadores de recicláveis, Revista Brasileira de Ciências Sociais, 23, 67.
- [4] Brasil, Instituto de Pesquisa Econômica Aplicada (2010): Pesquisa sobre o Pagamento por Serviços Ambientais Urbanos para Gestão de Resíduos, IPEA. Brasília.
- [5] Carmo, M. (2009): A Semântica ‘Negativa’ do Lixo Como Aspecto ‘Positivo’: Um Estudo de Caso Sobre uma Associação de Recicladores na Cidade do Rio de Janeiro, Brasil, Administração Pública e Gestão Social, Viçosa, 1. 2, 21-50, abr./jun.

- [6] Carmo, M. (2011): As Políticas de Estimulação ao Trabalho dos Catadores no Rio de Janeiro Como Estratégia Oriunda da Valorização dos Resíduos, Cadernos Gestão Pública e Cidadania (Notebooks Public Management and Citizenship): . 16, 58. São Paulo.
- [7] Christopher, M. (1997): Logística e gerenciamento da cadeia de suprimentos, Pioneira, São Paulo.
- [8] Ellram, L. (1991): Supply Chain management: the Industrial Organization Perspective, International Journal of Physical Distribution & Logistics Management, . 21, ° 1, 13-22.
- [9] Freires, F. (2000): Proposta de um modelo de gestão dos custos da cadeia de suprimentos. Dissertation (Master in Production Engineering): UFSC, Florianópolis.
- [10] Freires, F. (2007): A Influência dos Componentes Físicos e da Cooperação Entre os Actores Sobre o Desempenho de Sistemas Logísticos Inversos: o Caso dos Pneus-Resíduo. Thesis (Doctorate in Industrial Engineering and Management): Faculdade de Engenharia da Universidade do Porto, Porto.
- [11] Freires, F. and Guedes, A. (2006): A gestão do sistema logístico inverso para pneus-resíduo e sua relação com a eficácia e eficiência, XXVI Encontro Nacional de Engenharia de Produção, Fortaleza.
- [12] Freires, F. and Guedes, A. (2008): Power and Trust in Reverse Logistics Systems for Scruptires and its Impact on Performance, In. International Conference of the Production and Operations Management Society, Journal of Operations and Supply Chain Management 1(1): January-June, 57-65.
- [13] Freitas, L., Fonseca, I. (2012): Diagnóstico sobre Catadores de Resíduos Sólidos: Relatório de Pesquisa (Diagnosis of Collectors of Solid Waste: Research Report): IPEA, Brasília
- [14] Grimberg, E., Tuszal, L., Goldfarb, Y. (2004): Gestão Sustentável de Resíduos Sólidos e Inclusão Social: Estudo de caso sobre as Cooperativas das Centrais de Triagem do Programa Coleta Seletiva Solidária da Cidade de São Paulo (Sustainable Management of Solid Waste and Social Inclusion: Case Study on the Cooperatives of Screening Centers of the Joint Selective Collection Program of the City of São Paulo). ITCP/USP, São Paulo.
- [15] Hause, J. (1977): The measurement of concentrated industrial structure and the size distribution of firms, Annals of Economic and Social Measurement, 6(1): 73–107.
- [16] Henriques, R. (2004): Aproveitamento energético dos Resíduos Sólidos Urbanos: Uma abordagem Tecnológica, Dissertation (Master of Science in Energy Planning): UFRJ, Rio de Janeiro.
- [17] Kaplan, R. and Cooper, R. (1988): Custo e desempenho: administre seus custos para ser mais competitivo, Futura, São Paulo.
- [18] Krueger, A. (1974a): The Political Economy of Public Policy, Little, Brown, Boston.
- [19] Krueger, A. (1974b): The political economy of the rent-seeking society, American Economic Review, Princeton, NJ, 64, p. 291-303.
- [20] Magera, M. (2008): A Reciclagem dos Resíduos Sólidos Urbanos e o Uso das Cooperativas de Reciclagem. In.: http://www.fundace.org.br/cooperativismo/arquivos_pesquisa_ica_la_2008/043-magera.pdf. Latest access: 01 june 2009.
- [21] Martins, L., Peres, M. (2006): Economia Solidária: A Experiência da Cooperativa do Reciclador Solidário de Piracicaba, en Unimep, Anais. 4ª Mostra Acadêmica, 24 to 26 of October.
- [22] Primo, R. (2012a): Cooperativismo e Liberdade no Século XXI, CEALA, Salvador.
- [23] Primo, R. (2012b): Política Transversal de Regulação do Mercado Oligopsônico de Resíduos Sólidos nos Carnavais de Salvador de 2007 a 2012 por meio de um Complexo de Economia Solidária, Isegoria - Ação Coletiva em Revista, Universidade Federal de Viçosa – UFV, Viçosa, year 1, 1, 2.
- [24] Primo, R. (2014): Comércio Justo & Mercados Imperfeitos I: Fundamentação Teórica, Revista Dialética, 5, 5, 47.
- [25] Primo, R. (2015): Regulação de Logísticas Reversas nos Mercados Sub-Regionais de Reciclagem de Resíduos Sólidos: Enlace do Preço Justo Nacional aos Sistemas RFID. In. Brasil. Prêmio Celso Furtado 2015. The International Celso Furtado Center of Developed Politics: Ministério da Integração, Brasília.
- [26] Primo, R. and Rodríguez, J. (2015a): Project of Calculated Characterization of Monopsonic Degree into the Waste Solid Market in Large Metropolitan of Brazil, I Workshop Models and Methodologies to Support Decisions Applied to Health and Environment Management, 1(1): UFCE, Fortaleza.
- [27] Primo, R. and Rodríguez, J. (2015b): Between the Exploitation Oligopsonic and Fair Trade in Reverse Logistics of Solid Waste in Brazil: Mathematical Foundations, Transition Regulatory and Proposal of Policy in Three Steps. In.: Herrera, C. et al. (Ed.). Experiencias en la Modelación

- ción de la Toma de Decisiones en la Salud Humana, Medio Ambiente y Desarrollo Humano, Tomo 1, Rideca, UGR, Granada.
- [28] Primo, R. and Rodríguez, J. (2016a): Calculating Characterization of Monopsonic Degree in the Recycled Solid Waste Market in Metropolitan Regions of Brazil, Editorial Nova Publishers, New York.
- [29] Primo, R. and Rodríguez, J. (2016b): Globalización y Crisis Económica Neoliberal como Determinantes de Pobreza, Hambre y Desigualdad en América Latina, Congreso Internacional de Gestión Económica y Desarrollo, Asociación Nacional de Economistas y Contadores de Cuba, Habana.
- [30] Salvador (2015): Câmara Municipal. Irmandade de Salvador e Havana reforçada em sessão na Câmara. In: http://www.cms.ba.gov.br/noticia_int.aspx?id=9739. Latest access: 01 june 2016.
- [31] Posner, R. (1974): Theories of economic regulation, Bell Journal of Economics and Management Science, New York, 5, 3, 335-358.
- [32] Posner, R. (1975): The social costs of monopoly and regulation, Journal of Political Economy, Chicago, 83, 807-827.
- [33] Silva, M. et al (2014): Biodiesel in Brazil: A Market Analysis and Its Economic Effects, Journal of Agricultural Science; 6, 8, 160-178.
- [34] Silva, M. et al (2014): Family Farming and Biodiesel: The Difficulties of Socioeconomic Inclusion in the Northeast of Brazil, Journal of Agricultural Science; 6, 7, 231-252.
- [35] Silva, M. S.; Teixeira, F. L. C.; Torres, E. T. (2015): Desenvolvimento de um Instrumento Analítico para Diagnóstico de Competitividade da Cadeia Produtiva do Biodiesel no Estado da Bahia, Business and Management Review, 4, 9.
- [36] Soares, E. (2011): Estudo da caracterização gravimétrica e poder calorífico dos resíduos sólidos urbanos. Dissertation (Master in Civil Engineering): UFRJ, Rio de Janeiro.
- [37] Stigler, G. (1971): The theory of economic regulation, Bell Journal of Economics and Management Science, New York, 2, 1, 1-21.
- [38] Themelis, N. (2003): An Overview of the Global Waste-to-Energy Industry. In: Waste Management World, Jul-Aug 2003, 40-47.
- [39] Themelis, N. and Kaufman S. (2004): Waste in a Land of Plenty – Solid Waste Generation and Management in the US. In: Waste Management World, Sep-Oct 2004, 23-28.
- [40] Tolmasquin, M. (2003): Fontes Renováveis de Energia no Brasil. Interciência, Rio de Janeiro.

ANÁLISIS MULTIVARIANTE DE LA SOSTENIBILIDAD ECONÓMICA, AMBIENTAL Y SOCIAL, EN LAS EMPRESAS EUROPEAS (GRI)

Enrique Vicente-Galindo*; Belén Vaquero-Romero*; Purificación Vicente-Galindo*; Nathalia Tejedor-Flores*, Sergio Hernández-González** & Purificación Galindo-Villardón*, ***

*Universidad de Salamanca. España

**Universidad Veracruzana de Xalapa. Xalapa. México.

***ESPOL. Guayaquil. Ecuador

ABSTRACT

This research has analyzed the economic, environmental and social sustainability of 1987 reports published by European companies, through the GRI index, over the years 2011 and 2012 (GRI G3 and GRI G3.1), using multivariate methods designed for the inspection of dichotomous data, known as Biplot External Logistics, developed by Vincent-Villardón et al. (2006) and later extended by Demey et al. (2008), this technique is based on an algorithm that combines Principal Coordinates Analysis and Logistic Regression allowing simultaneous graphical representation of the companies and the sustainability characteristics studied, showing two gradients of sustainability that detect four clusters of companies. We have taken into account all European companies, regardless of size or economic activity and included in the classification of the United Nations: 34 countries, 1106 reports in 2011 and 983 in 2012. A comparison between European countries was conducted, differentiating between those who had both the highest GDP per capita and those countries defined as PIIGS that have been hardest hit by the effects of the financial crisis. We concluded that sustainability reports are of great importance for countries with high purchasing power, and for countries who have been most affected by the current economic crisis.

KEYWORDS: Sustainability, corporate social responsibility, Europe, GRI, External Logistic Biplot

RESUMEN

En esta investigación se ha analizado la sostenibilidad económica, ambiental y social de 1987 informes publicados por las empresas europeas, a través del índice GRI, a lo largo de los años 2011 y 2012 (GRI G3 y GRI G3.1), usando métodos multivariantes diseñados para la inspección de los datos dicotómicos, concretamente el Biplot Logístico Externo, desarrollado por Vicente-Villardón et al. (2006) y más tarde extendido por Demey et al. (2008), técnica que combina un Análisis de Coordenadas Principales y una Regresión Logística permitiendo una representación gráfica simultánea de las empresas y las características de sostenibilidad estudiadas; poniendo de manifiesto dos claros gradientes de sostenibilidad que permiten detectar a su vez 4 clústeres de empresas. Se han tenido en cuenta todas las empresas, independientemente de su tamaño o actividad económica y se ha considerado como pertenecientes a Europa todos los países incluidos en la clasificación de las Naciones Unidas: 34 países, 1106 informes en 2011 y 983 en 2012. Se realizó una comparación entre los países europeos, diferenciando entre aquellos que tenían el mayor PIB per cápita y aquellos países definidos como PIIGS que han sido los más afectados por los efectos de la crisis financiera. Se concluye que los informes de sostenibilidad son de gran importancia tanto para los países con alto poder adquisitivo como para los países más afectados por la actual crisis económica.

1. INTRODUCCIÓN

La búsqueda del equilibrio entre las necesidades del hombre y los recursos disponibles ha sido constante; sin embargo, es en 1987 cuando se establece, a nivel mundial, la expresión “desarrollo sostenible”. El *Informe Brundtland*, publicado por la Comisión Mundial sobre el Medio Ambiente y el Desarrollo (CMMAD) estableció la definición más actual del término *desarrollo sostenible*: “Como aquel que satisface las necesidades de la generación actual sin comprometer la capacidad de las generaciones futuras para satisfacer sus propias necesidades”. En el 2012 se celebró en Río de Janeiro la Conferencia de las Naciones Unidas sobre el Desarrollo Sostenible Río+20, que tenía entre sus objetivos, el establecimiento de un marco institucional para el desarrollo sostenible (UNCSD, 2012). Sin embargo, la humanidad está lejos aún de encaminarse hacia la sostenibilidad (Waas, Verbruggen y Wright, 2011).

Desde el punto de vista empresarial, el modelo de gestión tradicional, basado únicamente en la maximización del valor para los accionistas, se vio obligado a adaptarse a unas nuevas exigencias que incluían la parte social y ambiental en la toma de decisiones. A partir de los años 90 se empezó a desarrollar en Europa el concepto de Responsabilidad Social Corporativa (RSC), el cual se puede definir como “el

reconocimiento e integración en sus operaciones por parte de las empresas, de las preocupaciones sociales y medioambientales, dando lugar a prácticas empresariales que satisfagan dichas preocupaciones y configuren las relaciones con sus interlocutores” (Valor Martínez y De la Cuesta González, 2003). Esto contribuyó a que las compañías comenzaran a publicar información relacionada con sus actividades, favoreciendo el posterior desarrollo de los llamados *Informes de Sostenibilidad*. Según Moneva (2007), los *Informes de Sostenibilidad* serán aquellos que pretenden comunicar a todos los grupos de interés, tanto los compromisos de la organización respecto al desarrollo sostenible, como sus logros, detallando los resultados de sus actuaciones en los ámbitos económico, social y medioambiental.

En 1997 aparece el Global Reporting Initiative (GRI): En año 2004 se convirtió en una entidad independiente y a día de hoy su marco para la elaboración de informes de sostenibilidad goza de gran credibilidad a nivel mundial. Desde que en el año 2000 el GRI emitiera su primera guía para la elaboración de memorias de sostenibilidad, han sido varias las que le han sucedido siendo la G4 la cuarta y más reciente versión de todas ellas, presentada en mayo de 2013. No obstante, debido al bajo número de memorias publicadas en el momento de la elaboración de este trabajo, nuestro estudio se centrará en las versiones G3 y G3.1.

El Marco GRI para la elaboración de memorias de sostenibilidad ha sido diseñado para ser utilizado por las organizaciones con independencia de su tamaño, sector o localización, y se espera que sirva como marco generalmente aceptado para informar acerca del desempeño económico, ambiental y social (GRI, 2006).

Las organizaciones, una vez concluida la memoria, deben indicar en qué nivel han aplicado el Marco de elaboración de memorias del GRI a través del sistema de “niveles de aplicación del GRI”. Existen tres niveles de aplicación que permiten cumplir las necesidades de las organizaciones informantes. Los niveles son, ordenados de menor a mayor: C, B y A. Asimismo, una organización podrá añadir a estas calificaciones un “plus” (+) en cada nivel (C+, B+, A+), en el caso de que se haya utilizado verificación externa.

El objetivo de esta investigación es realizar un análisis de la sostenibilidad económica, ambiental y social en empresas europeas, a través del índice GRI (Global Reporting Initiative), utilizando para ello métodos multivariantes especialmente diseñados para la inspección de datos dicotómicos.

2. MATERIAL Y MÉTODOS

Para llevar a cabo esta investigación se ha procedido a analizar los informes de sostenibilidad publicados por empresas localizadas en países europeos a lo largo de los años 2011 y 2012. Para ello, hemos utilizado la herramienta de búsqueda disponible en la base de datos de GRI (<http://database.globalreporting.org/search>), donde se ha procedido a la búsqueda de todas aquellas empresas que hubieran publicado informes de sostenibilidad basados en las directrices de las guías G3 y G3.1. Se han tenido en cuenta todas las empresas independientemente del tamaño o del sector de actividad de cada una de ellas. Respecto al área geográfica de aplicación, hemos considerado como pertenecientes a Europa todos aquellos países incluidos en la clasificación que las Naciones Unidas realizan por regiones (United Nations Statistics Division, 2013):

Los 34 países europeos que a lo largo del periodo de tiempo mencionado publicaron informes de sostenibilidad, así como el número de informes publicados por cada uno de ellos, se detallan en la Tabla 1.

La información resultante ha sido organizada en una matriz (X) de datos binarios IxJ donde las I filas corresponden a los 1987 informes publicados por las distintas empresas analizadas y las J columnas corresponden a los 55 indicadores principales GRI incluidos en nuestro estudio, los cuales hacen referencia a la *dimensión económica* (EC1, EC2, EC3, EC4, EC6, EC7, EC8), la *dimensión ambiental* (EN1, EN2, EN3, EN4, EN8, EN11, EN12, EN16, EN17, EN19, EN20, EN21, EN22, EN23, EN26, EN27, EN28), la *dimensión laboral* (LA1, LA2, LA4, LA5, LA7, LA8, LA10, LA13, LA14, LA15), la *dimensión de derechos humanos* (HR1, HR2, HR3, HR4, HR5, HR6, HR7, HR10, HR11), la *dimensión social* (SO1, SO2, SO3, SO4, SO5, SO8, SO9, SO10) y la *dimensión de responsabilidad sobre productos* (PR1, PR3, PR6, PR9): El detalle de los indicadores puede consultarse en Guía para la elaboración de Memorias de Sostenibilidad, Versión 3.1 (GRI, 2011): La descripción de los índices se encuentra en las páginas 29 a la 43.

Tabla 1. Número de informes de sostenibilidad GRI por países. Años 2011 y 2012

País	2011	2012	Nº Total Informes
Albania	1	0	1
Alemania	118	96	214
Andorra	2	2	4
Austria	46	49	95
Bélgica	25	25	50
Bulgaria	3	3	6
Croacia	7	6	13
Dinamarca	17	18	35
Eslovaquia	1	2	3
Eslovenia	4	6	10
España	172	89	261
Finlandia	59	70	129
Francia	26	24	50
Grecia	28	17	45
Hungría	27	19	46
Irlanda	5	7	12
Islandia	1	2	3
Italia	57	46	103
Letonia	1	1	2
Luxemburgo	5	7	12
Noruega	22	18	40
Países Bajos	88	77	165
Polonia	17	25	42
Portugal	37	15	52
Reino Unido	53	63	116
República Checa	3	1	4
Rep. de Moldavia	0	1	1
Rusia	56	51	107
Serbia	4	6	10
Suecia	108	109	217
Suiza	74	85	159
Turquía	31	30	61
Ucrania	4	8	12
TOTAL	1106	983	2089

Nota: Elaborado por Belén Vaquero Romero a través de la plataforma del GRI. Países con menos de 15 informes no formaran parte del análisis multivariante.

Tanto en el año 2011 como en el 2012 los indicadores más publicados se encontraban dentro de los económicos (EC1), los medioambientales (EN3 y EN16) y los laborales (LA1, LA7 y LA13), con un porcentaje del 80% o superior.

Los sectores de actividad seleccionados para llevar a cabo nuestro estudio se corresponden con los establecidos por NACE (Parlamento Europeo, 2006), los cuales se han agrupado así, (Tabla 2):

Tabla 2. Identificador de los sectores

IDENTIFICADOR DEL GRUPO	SECTOR (NACE)
Ay B	Sector Primario (SP)
C	Sector Industrial (SI)
D y E	Agua y Energía (AE)
F	Construcción (CN)
Desde la G hasta la N	Sector Servicios (SS)
O y Q	Administración Pública y Servicio Sociales (AS)
P	Educación (ED)
Desde la R hasta la U	Otros (OT)

Nota: Parlamento Europeo (2006): Nomenclatura estadística de actividades económicas NACE.

2.1. Métodos Estadísticos

Las técnicas tradicionales de reducción de dimensiones, como pueden ser el Análisis de Componentes Principales o el Análisis Factorial, si bien son ampliamente utilizadas, no son apropiadas para nuestros datos al estar trabajando con variables binarias. Para nuestra investigación utilizaremos la técnica conocida como **Biplot Logístico Externo**, desarrollada por Demey, Vicente-Villardón, Galindo y Zambrano (2008), basada en un algoritmo que combina el Análisis de Coordenadas Principales (ACoP) y la Regresión Logística (RL): El algoritmo permite una representación gráfica simultánea de las empresas y de las características de sostenibilidad que se estudian, haciendo la interpretación más sencilla y permitiendo una exploración visual más intuitiva del conjunto de datos.

Sea un matriz de datos binarios X , donde la presencia del atributo se representa con 1's y la ausencia se representa con 0's, y donde las filas corresponden a las n empresas europeas consideradas y las columnas a los p caracteres binarios (indicadores GRI): Sea, además, $S = (s_{ij})$ la matriz conteniendo las similitudes entre las empresas obtenidas a partir de la matriz de datos X , y sea $\Delta = (\delta_{ij})$ la matriz de disimilitudes donde $\delta_{ij} = 1 - s_{ij}$.

El algoritmo comienza con un ACoP, que trata de construir una configuración de n puntos en un espacio euclídeo, de tal forma que la distancia entre dos puntos cualesquiera de la configuración aproxime, tan cerca como sea posible, la disimilitud (δ_{ij}) entre las empresas que, en nuestro caso, son representadas por estos puntos. De manera gráfica, resumimos:

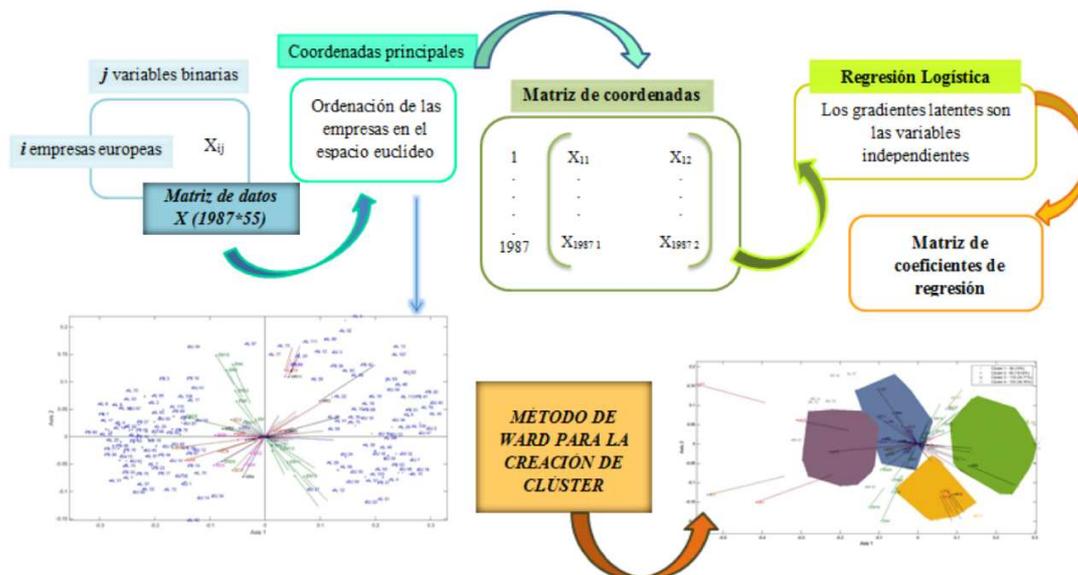


Figura 1. Pasos del Biplot Logístico Externo. Elaborado por Belén Vaquero Romero.

El objetivo entonces será encontrar una configuración Y en espacio euclídeo de baja dimensión \mathbb{R}^k cuya matriz de distancias entre puntos D esté tan cerca como sea posible a Δ . Cuando la disimilitud medida es euclídea, es posible encontrar una configuración exacta en $n-1$ dimensiones. Además, se puede obtener una aproximación de la dimensión más reducida proyectando en las primeras k coordenadas principales (normalmente $k=2$):

En ACoP es sabido que la proporción de la varianza total explicada por las k dimensiones puede ser considerada como una media de bondad de ajuste de los n puntos en la representación gráfica. Sin embargo, una buena bondad de ajuste global no implica que todas las empresas tengan la misma calidad de representación y, por tanto, la interpretación de las posiciones de todos los puntos en el diagrama no es igualmente fiable. Se considera que un individuo está bien representado cuando la mayor parte de su información, medida a través de la variabilidad, se contabiliza en la dimensión reducida. Como la representación está centrada en el origen, la variabilidad de cada individuo se mide por el cuadrado de su distancia al centro, de forma que la calidad de la representación puede ser medida por el ratio entre el

cuadrado de la distancia en la dimensión reducida y el cuadrado de la distancia en el espacio completo. Esto es:

$$CR_i^k = \frac{\sum_{i=1}^k y_{i1}^2}{\sum_{j=1}^k y_{ij}^2} \times 100\%$$

Donde y_{ij} representa las coordenadas principales del individuo i en la j -ésima dimensión. Geométricamente, es el cuadrado del coseno del ángulo entre el vector en el espacio completo y su proyección en el espacio de representación.

Al contrario que el Análisis de Componentes Principales en su versión Biplot, donde los nuevos ejes pueden ser interpretados en términos de las variables originales, en ACoP los ejes no tienen significado directo. Por tanto, no es posible interpretar la relación entre las empresas y los indicadores GRI. No obstante, se puede demostrar que las configuraciones del Análisis de Componentes Principales se pueden también obtener aplicando ACoP a la matriz de distancias euclídeas. Para buscar las variables asociadas a la ordenación obtenida en el ACoP, podemos buscar las direcciones en el diagrama de ordenación que mejor predicen la probabilidad de presencia de cada indicador GRI. De esta forma, definimos $\pi_{ij} = E(x_{ij})$ como la probabilidad esperada de que el indicador j esté presente en una empresa con coordenadas y_{is} ($i=1,\dots,n$; $s=1,\dots,k$) en el diagrama de ordenación. Entonces,

$$\pi_{ij} = \frac{e^{b_{j0} + \sum_{s=1}^k b_{js}y_{is}}}{1 + e^{b_{j0} + \sum_{s=1}^k b_{js}y_{is}}}$$

Donde b_{js} ($j = 1, \dots, p$) son los coeficientes de RL que corresponden con la j variable en la s -ésima dimensión. El modelo es un modelo lineal generalizado conteniendo el logit como una función enlace.

$$\text{logit}(\pi_{ij}) = \log\left(\frac{\pi_{ij}}{1 - \pi_{ij}}\right) = b_{j0} + \sum_{s=1}^k b_{js}y_{is} = b_{j0} + y_i^T b_j$$

Donde $y_i = (y_{i1}, \dots, y_{ik})^T$ y $b_j = (b_{j1}, \dots, b_{jk})^T$, y las y 's y b 's definen al biplot en una escala logit. A esto se le conoce como Biplot Logístico Externo porque las coordenadas de las empresas son calculadas con un procedimiento externo (ACoP): Dado que las y 's son conocidas del ACoP, obtener las b 's es equivalente a realizar una RL usando la columna j -ésima de la matriz X como variable respuesta y las columnas de y como variables regresoras. La ecuación de regresión predecirá la probabilidad de que un indicador GRI esté presente en una empresa. Geométricamente, las y 's pueden ser representadas como puntos en un espacio de dimensión reducida y las b 's como vectores que muestran las direcciones que mejor predicen la probabilidad de presencia de cada indicador π_{j} .

Para cada indicador considerado en el estudio, el diagrama de ordenación puede ser dividido en dos regiones que predicen, respectivamente, la presencia/ausencia de la característica. Las dos regiones están separadas por una línea perpendicular al vector que representa a la variable y corta al vector en el punto 0,5. Si la empresa que estamos analizando está en la dirección que marca el extremo del vector, esa característica estará presente. La calidad de representación de cada variable viene medida por el p-valor en el modelo de regresión logística.

Con el análisis también tratamos de encontrar países con perfiles de sostenibilidad similares (agrupaciones): El Análisis Clúster se puede aplicar utilizando la matriz de distancias inicial o la matriz de distancias euclídeas ajustadas D obtenida a partir del ACoP. La partición obtenida es representada en la ordenación del ACoP utilizando la envolvente convexa de los puntos pertenecientes a cada clúster. En nuestro caso, para la creación de clústers utilizaremos el método de Ward de varianza mínima, basado en el procedimiento de aglomeración de clúster jerárquico donde el criterio para la elección del par de clúster a mezclar en cada paso está basado en el valor óptimo de una función objetivo, el error de la suma de los cuadrados, con la ordenación de las puntuaciones del Biplot Logístico (Vicente Galindo, Vaz y de Noronha, 2015).

En términos del análisis del gráfico, destacamos las siguientes reglas de interpretación:

- Las distancias entre los puntos que representan a las empresas en el diagrama de dispersión están inversamente relacionadas a las similitudes de sus perfiles. De esta forma, empresas que se encuentren cerca tendrán similares características.

- El ángulo entre los vectores que representan las variables indica el grado de asociación entre variables: los ángulos agudos indican que las variables están estrechamente relacionadas, es decir, cuando una empresa presenta una de las características también presenta la otra y viceversa.
- Los ángulos entre los vectores representando las variables y el eje factorial, estiman el grado de relación entre la variable y la dimensión latente.
- Las proyecciones de las empresas en el vector que representa a las variables, estiman la probabilidad esperada de las características de esa empresa dada su combinación de características.
- La longitud del vector que representa cada variable indica el poder de discriminación de la variable en la separación de las empresas. Los vectores más cortos tendrán el mayor poder de discriminación.

El programa estadístico utilizado para implementar esta técnica, elaborado por Vicente Villardón, se conoce como MultBiplot (Vicente-Villardón, 2015) y está disponible para su descarga en la página web: <http://biplot.dep.usal.es/classicalbiplot/>

3. RESULTADOS

3.1. Análisis Exploratorio

Analizando los datos obtenidos en el estudio, observamos que el número de publicaciones es muy similar en ambos años y que, además, la mayoría de las compañías estudiadas tienen un tamaño grande, son privadas y, casi la mitad de ellas, pertenecen al sector servicios. El porcentaje entre aquellas que publican de acuerdo a las directrices GRI 3 es ligeramente superior al de aquellas que publican de acuerdo a GRI 3.1, destacando aquellas compañías que presentan niveles de aplicación de A+, B y C.

Para estudiar el comportamiento en materia de sostenibilidad de los países, diferenciaremos entre los que en el periodo analizado tenían el mayor PIB per cápita de Europa, eligiendo como población objetivo las empresas de aquellos países que presenten al menos 150 publicaciones entre los años 2011 y 2012, y los países denominados PIIGS (Portugal, Italia, Irlanda, Grecia y España) que se han visto más afectados por los efectos de la crisis financiera. Es importante indicar que Irlanda quedará excluida del estudio al tener un número de publicaciones inferior a 15.

De acuerdo a los datos publicados por el Banco Mundial, organismo dependiente de las Naciones Unidas (Banco Mundial, 2015), los países con mayor PIB per cápita en el periodo objeto de estudio, que a su vez presentan un mayor número de informes de sostenibilidad, son Alemania con 118 informes en 2011 y 96 en 2012, Países bajos con 88 y 77 respectivamente, Suecia 108 en 2011 y 109 en 2012 y Suiza 74 en 2011 y 85 en 2012. Los países denominados PIIGS, afectados de manera muy significativa por la crisis, son: España con 172 reportes en 2011 y 89 en 2012, Grecia con 28 en 2011 y 17 en 2012, Italia con 57 en 2011 y 46 en 2012 y Portugal con 37 en 2011 y 15 en 2012.

En el año 2011, dentro de las empresas situadas en países con mayor PIB per cápita de Europa, los indicadores más publicados corresponden a la parte *económica*, EC1 (84,2%); a la parte *medioambiental*, EN16 (81,7%); y a la parte *laboral*, LA1 (89,4%); Lo mismo sucede en el año 2012 con porcentajes ligeramente superiores. Respecto a las empresas localizadas en aquellos países denominados PIIGS, los más afectados por la crisis, hemos observado que, en el año 2011, los indicadores más publicados corresponden a la parte *económica*, EC1 (92,9%); a la parte *medioambiental*, EN3 (86,1%); y a la parte *laboral*, LA1 (91,8%); En 2012, los indicadores más publicados corresponden a la parte *económica*, EC1 (95,8%); a la parte *medioambiental*, EN22 (90,4%); a la parte *laboral*, LA1 (95,8%), y LA7 (95,8%); a la parte *social*, SO1 (82,0%); y a la parte de *responsabilidad sobre productos*, PR3 (80,8%): Los indicadores exclusivos de la versión G3.1 son de nuevo los menos publicados: LA15 (46,1%), HR10 (41,9%), y SO10 (44,9%):

3.2. Gradientes de Sostenibilidad

Como ya se ha visto, el método a aplicar para obtener los principales gradientes de sostenibilidad y su relación con las características observadas, así como su representación gráfica, será el Biplot Logístico Externo (Demey et al., 2008):

Comenzaremos estudiando el caso de las empresas que se encuentran en **países con mayor PIB per cápita de Europa**, en el **año 2011**. De acuerdo a los resultados obtenidos, el primer plano principal para una solución bidimensional recoge el 35,20% de la variabilidad. El primer valor propio (12,69) es

significativamente mayor que el segundo (1,52), lo que significa que incluso en el caso de que se consideren los dos gradientes de sostenibilidad, la primera dimensión (horizontal) recoge la mayor parte de la información. La bondad de ajuste global, como porcentaje de clasificación correcta en el biplot, es 82,89%, por lo que con una solución de dos dimensiones será suficiente para explorar las características principales de los datos.

La solución biplot obtenida se muestra en la Figura 2. Para crear clústeres de empresas con un perfil similar de sostenibilidad usamos el método jerárquico de Ward sobre las puntuaciones del Biplot Logístico.

Todas las variables están significativamente relacionadas con los ejes de ordenación del primer plano principal ya que todas tienen un p-valor menor que 0,05. Las variables con un R^2 mayor de 0,80 pueden ser consideradas las más relevantes para la interpretación de la ordenación de las empresas en el diagrama de dispersión. En nuestro caso, estas variables son: HR11 (0,87) y SO9 (0,80): El gradiente asociado al eje 1 está principalmente relacionado con las variables EC3, EC6, EC8, (Económicas), EN26, (Medioambientales), LA4, LA7, LA8, LA10, LA15, (Laborales), HR1, HR4, HR5, HR6, HR11, (Recursos Humanos), SO4, (Sociales) y PR1, PR3, (Responsabilidad de Productos): El gradiente asociado al eje 2 está relacionado con las variables EN1, EN2, EN3, EN4, EN16, EN20, EN22, EN27, (Medioambientales), LA2, LA14, (Laborales), HR3, HR7, HR10, (Recursos Humanos), SO3, SO5, SO8, SO9, (Sociales) y PR6 (Responsabilidad de Productos):

Existen, pues, dos gradientes principales. El primero está relacionado con el desempeño económico, la presencia en el mercado y los impactos económicos indirectos, con los productos y servicios, con el empleo, con las relaciones empresa-trabajadores, con la salud y seguridad laboral, con la formación y evaluación, con las prácticas de inversión y abastecimiento, la no discriminación, la libertad de asociación, el trabajo infantil y las medidas correctivas, con la corrupción, con la salud y seguridad del cliente, y con el etiquetado de productos y servicios. El segundo está relacionado con los materiales utilizados, con el consumo de energía, con las emisiones, vertidos y residuos, con los productos y servicios, con las prácticas de inversión y abastecimiento, con el trabajo forzado y la evaluación de los derechos humanos, con la corrupción y la política pública, con el cumplimiento normativo y el impacto en comunidades locales, y con las comunicaciones de marketing.

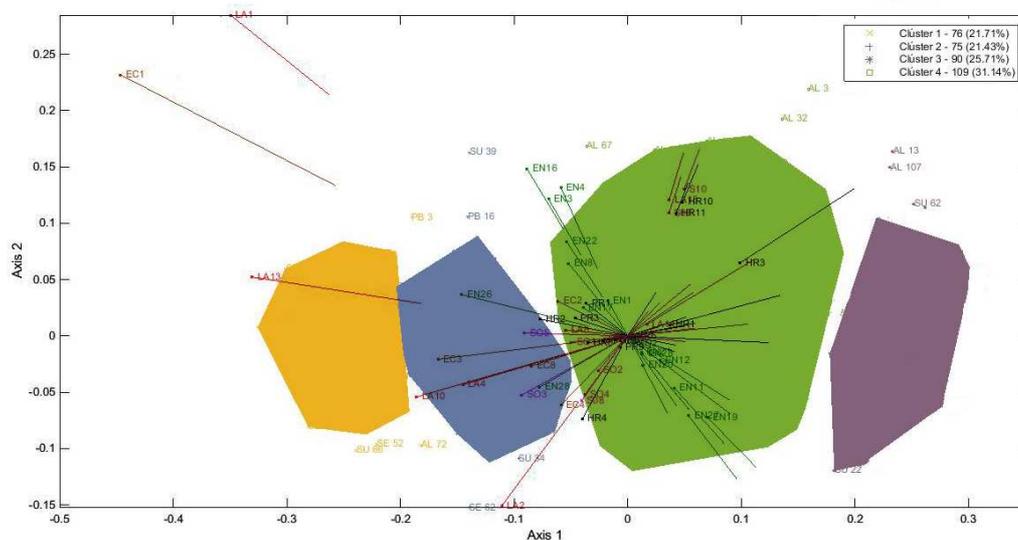


Figura 2. Biplot Logístico Externo de dos dimensiones con clústeres, basado en los datos de sostenibilidad de empresas de países con mayor PIB per cápita. Año 2011. Elaborado por Belén Vaquero Romero.

En la Figura 2, las empresas en cada clúster se han coloreado respectivamente y se ha dibujado un polígono conteniendo todas las empresas. Cuatro clúster han sido descritos y representados en el mapa euclídeo de esta figura.

Hay ciertos indicadores que presentan una bondad de ajuste poco elevada porque su variabilidad no es alta (se puede deber a que sea un indicador publicado por la mayoría de las empresas) o porque necesitamos más dimensiones para describir a todo el conjunto de los indicadores. No obstante, una solución bidimensional es suficiente para recoger la información más relevante:

- El clúster 3, representado en color morado, que contiene 90 empresas (25,71%), es el clúster formado por las empresas que publican un mayor número de índices de sostenibilidad, por lo que podemos decir que contiene a las empresas con el mayor nivel de contribución a la sostenibilidad de sus países. Cabe destacar que el 100% de las empresas del clúster 3 publican los índices EC1, EN16, EN22, LA1 y LA13. Más del 90% publican el resto de índices, con excepción de: EN11, EN19 y LA14, publicados por más del 80%; EN27, publicado por más del 75%; HR3, publicado por el 63,33%; y LA15, HR10, HR11, SO9 y SO10, publicados por el 30% o menos de las empresas. Debe tenerse en cuenta que los índices LA15, HR10, HR11, SO9 y SO10 fueron incorporados en la guía G3.1, por lo que están ausentes en todas las publicaciones anteriores.
- El clúster 4, representado en color verde, contiene a las 109 empresas (31,14%) que presentan una sostenibilidad media-alta, si bien ningún indicador ha sido publicado por el 100% de ellas. No obstante, más del 90% publican los indicadores EC1, EN3, EN16, LA1 y LA13. Además, más del 80% publican los indicadores EC3, EN4, EN22, EN26, LA2, LA7, SO3, SO4 y SO5. Entre el 70% y el 79% publican los indicadores EC2, EC8, EN8, EN28, LA4, LA10, HR2, HR4, HR5, HR6, HR7 y SO8. Entre el 50% y el 69% publican los indicadores EC4, EC6, EC7, EN1, EN2, EN17, LA5, LA8, LA14, SO1, SO2, PR1, PR3, PR6 y PR9. Menos del 50% publican los indicadores EN11, EN12, EN19, EN20, EN21, EN23, EN27, LA15, HR1, HR3, HR10, HR11, SO9 y SO10.
- El clúster 2, representado en color azul, contiene 75 empresas (21,43%) con sostenibilidad media-baja, donde destaca que todas ellas han publicado el índice LA1. Además, más del 80% publican los indicadores EC1, EN16, LA7 y LA13. Entre el 70 y el 79% publican los indicadores EN3, EN4 y LA2. Entre el 50% y el 69% publican los índices EC3, EN8, EN22, EN26, LA10, HR4 y SO8. El resto de índices son publicados por menos del 50% de las empresas.
- El clúster 1, representado en color amarillo, contiene a las 76 empresas (21,71%) con menor contribución a la sostenibilidad. Únicamente los indicadores EC1 y LA1 están publicados por más del 80% de las empresas. Además, observamos que, de los 55 indicadores analizados, 41 son publicados por menos del 30% de las empresas.

Analizaremos a continuación el caso de estos mismo países, en el **año 2012**.

La bondad de ajuste global como porcentaje de clasificación correcta en el biplot es 82,54%, por lo que con una solución de dos dimensiones será suficiente para explorar las características principales de los datos.

La solución biplot obtenida se muestra en la Figura 3.

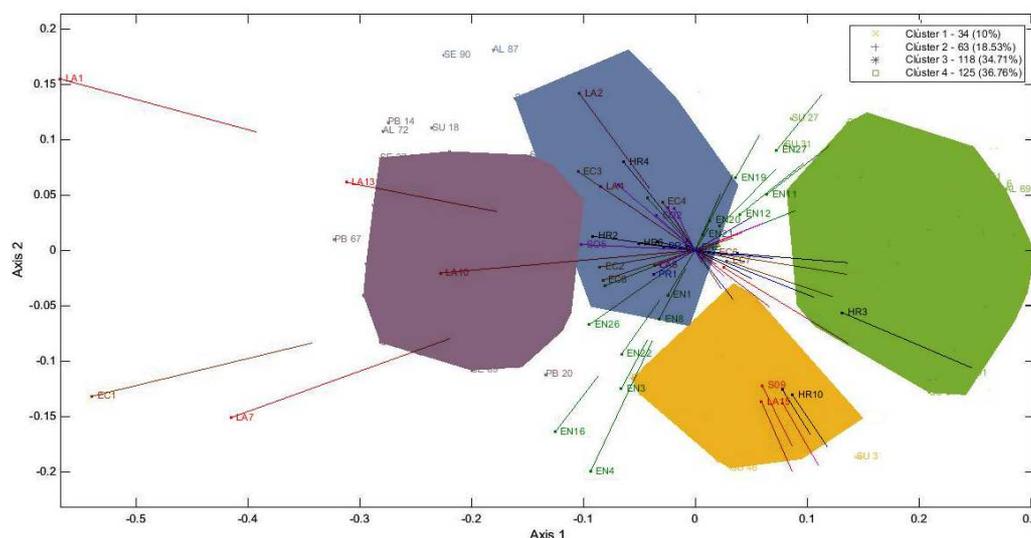


Figura 3. Biplot Logístico Externo de dos dimensiones con clústeres, basado en los datos de sostenibilidad de empresas de países con mayor PIB per cápita. Año 2012. Elaborado por Belén Vaquero Romero.

Cuatro clúster han sido descritos también y representados en el mapa euclídeo de esta figura. Todas las variables están significativamente relacionadas con la ordenación observada de las empresas ya que todas

tienen un p-valor menor que 0,05. Las variables con mayor R^2 serán las más relevantes para la interpretación de la ordenación de las empresas en el diagrama de dispersión. En nuestro caso, estas variables son: EN19 (0,74) y HR5 (0,74):

El gradiente asociado al eje 1 está principalmente relacionado con las variables EC1, EC2, EC6, EC7, EC8 (Económicas), EN17 (Medioambientales), LA1, LA7, LA8, LA10, LA13 (Laborales), HR1, HR2, HR3, HR6, HR7 (Recursos Humanos), SO1, SO5 (Sociales), PR3, PR6 (Responsabilidad de Productos): El gradiente asociado al eje 2 está relacionado con las variables EC4 (Económicas), EN1, EN2, EN3, EN4, EN8, EN16, EN19, EN20, EN21, EN22 (Medioambientales), LA2, LA15 (Laborales), HR10, HR11 (Recursos Humanos), SO4, SO8, SO9, SO10 (Sociales):

También se han encontrado dos gradientes principales. El primero está relacionado con el desempeño económico, la presencia en el mercado y los impactos económicos indirectos, con las emisiones, vertidos y residuos, con el empleo, con la salud y la seguridad laboral, con la formación y evaluación y con la diversidad e igualdad de oportunidades, con las prácticas de inversión y abastecimiento, con el trabajo infantil y los trabajos forzados, con el impacto en las comunidades locales y las políticas públicas, con el etiquetado de productos y servicios y con las comunicaciones de marketing. El segundo está relacionado con el desempeño económico, con los materiales utilizados y el consumo de energía, con las emisiones, vertidos y residuos, con el empleo, con la evaluación y las medidas correctivas en el campo de los derechos humanos, con la corrupción, el cumplimiento normativo y el impacto en las comunidades locales. Los resultados son similares pero hay un cambio en la configuración de los gradientes.

La información obtenida con una solución bidimensional, es la siguiente:

- El clúster 4, representado en color verde, está formado por las 125 empresas (36,76%) con mayor nivel de sostenibilidad. Cabe destacar que el 100% de ellas publican el índice EN16 y más del 90% publican los índices EC1, EC2, EC3, EC8, EN1, EN3, EN4, EN8, EN22, EN26, EN28, LA1, LA2, LA4, LA7, LA8, LA13, HR2, HR4, HR5, HR6, HR7, SO1, SO2, SO3, SO4, SO5 y PR1. Entre el 80% y el 89% publican los índices EC4, EC6, EC7, EN2, EN12, EN17, EN21, EN23, LA5, LA10, HR1, SO8, PR3, PR6 y PR9. Entre el 70% y el 79% publican los índices EN11, EN20 y LA14. Entre el 50% y el 69% publican los índices EN19, EN27, HR3. Menos del 35% publican los indicadores LA15, HR10, HR11, SO9 y SO10. No obstante, debe tenerse en cuenta que estos son los índices que se añadieron en la guía G3.1.
- El clúster 1, representado en color amarillo, contiene a las 34 empresas (10%) que presentan sostenibilidad media-alta. Destacamos que el 100% de las empresas publican el índice LA1 y más del 90% publican los índices EC1, EC3, EN16, LA2, LA13, HR4 y SO4. Entre el 80% y el 89% publican los índices EC4, EN28, LA7, HR2, SO2, SO3, SO5, SO8 y PR9. Entre el 70% y el 79% publican los índices EC2, EN3, EN4, LA4, LA5, LA10, HR5, HR6, HR7 y PR3. Entre el 50% y el 69% publican los índices EC6, EC7, EC8, EN22, EN26, LA8, LA14, SO1, PR1 y PR6. Menos del 50% publican el resto de indicadores.
- El clúster 2, representado en color azul, contiene 63 empresas (18,53%) con sostenibilidad media-baja, de las cuales más de un 90% publican los índices EC1, EN3, EN4, EN16 y LA1. Entre el 80% y el 89% publican los índices EN8, EN22, LA7 y LA13. Entre el 70% y el 79% publican el índice EN26 y entre el 50% y el 69% publican los índices EC2, EC3, EC8, EN1, EN2, EN28, LA2, LA4, LA10, HR2, HR4, SO1, SO3, SO5, PR1 y PR3. Menos del 50% de las empresas publican el resto de los índices GRI.
- El clúster 3, representado en color morado, contiene las 118 empresas (34,71%) con menor nivel de sostenibilidad. Únicamente los índices EC1, EN4, EN16, LA1, LA7 y LA13 están publicados por más del 70% de las empresas. El resto de índices están publicados por un porcentaje de empresas significativamente menor.

Alemania es el país que en ambos años presenta un mayor número de compañías en el clúster de máxima sostenibilidad, diferenciándose de manera significativa con el resto de los países analizados. De la misma forma, si tenemos en cuenta el porcentaje de empresas que, dentro de cada país, se encuentran dentro de dicho clúster, queda patente la ventaja de Alemania sobre el resto de países, concluyendo por tanto que, en base a nuestros datos, Alemania es el país más comprometido en materia de sostenibilidad. (Ver Anexos Tabla 3 y 4):

Además, nuestro estudio nos destaca que las empresas más sostenibles son, en su mayoría, de gran tamaño, privadas y pertenecientes al sector servicios. Dentro de este grupo de empresas hay una clara tendencia a la utilización de las directrices de la guía G3, haciendo uso de la comprobación GRI y del nivel de aplicación A+ en la publicación de sus informes.

Respecto al clúster con menor sostenibilidad, tanto en el año 2011 como en el 2012, destacamos que está formado en su mayoría por empresas de Suecia, que también presenta unos niveles muy altos de inclusión en el mencionado clúster dentro del propio país. Se hace por tanto evidente que, a pesar del gran número de informes que las empresas suecas publican cada año, su compromiso con la sostenibilidad tiene aún un largo camino que recorrer. (Ver Anexos Tablas 5 y 6):

De las empresas menos sostenibles la mayoría son de gran tamaño, privadas y pertenecientes al sector servicios. Se tiende nuevamente a la utilización de las directrices de la guía G3, haciendo uso en este caso de la verificación externa y de los niveles de aplicación C y C+ en la publicación de sus informes.

Una vez visto el comportamiento en materia de sostenibilidad de los países con mayor PIB per cápita de Europa (Ver Anexos Tablas 7, 8, 9 y 10), nos centraremos en comparar el país resultante como más sostenible, Alemania, con los **países denominados PIIGS**, fuertemente afectados por la crisis económica, con el objetivo de analizar si el poder adquisitivo que presentan los países tiene alguna relación en su desempeño a nivel de sostenibilidad. Comenzaremos el estudio por el **año 2011**.

De acuerdo a nuestros resultados, el primer plano principal para una solución bidimensional recoge el 36,88% de la variabilidad. El primer valor propio (13,52) es significativamente mayor que el segundo (2,32), lo que significa que incluso en el caso de que se consideren los dos gradientes de sostenibilidad, la primera dimensión (horizontal) recoge la mayor parte de la información. El ajuste de bondad global como porcentaje de clasificación correcta en el biplot es 84,39%, por lo que con una solución de dos dimensiones será suficiente para explorar las características principales de los datos.

Los resultados obtenidos acerca de la bondad de ajuste de los indicadores GRI son los siguientes:

Todas las variables están significativamente relacionadas con la ordenación observada de las empresas y aquellas con un R^2 mayor de 0.80 pueden ser consideradas las más relevantes para la interpretación de la ordenación en el diagrama de dispersión. En nuestro caso, estas variables son: HR10 (0,88), HR11 (0,92), SO9 (0,87) y SO10 (0,84):

El gradiente asociado al eje 1 está principalmente relacionado con las variables EC3, EC4, EC6, EC8, (Económicas), EN28 (Medioambientales), LA1, LA2, LA4, LA7, LA8, LA14, (Laborales), HR1, HR2, HR4, HR5, HR6, HR7, (Recursos Humanos), SO1, SO2, SO3, SO4, SO5, (Sociales): El gradiente asociado al eje 2 está relacionado con las variables EN2, EN3, EN16, EN27, (Medioambientales), LA15, (Laborales), HR10, HR11, (Recursos Humanos), SO9, SO10, (Sociales):

Hemos encontrado dos gradientes principales: El primero está relacionado con el desempeño económico, la presencia en el mercado y los impactos económicos indirectos, con el cumplimiento normativo, el empleo, con la salud y la seguridad laboral, con la igualdad de oportunidades, con las prácticas de inversión, con la libertad de asociación, con el trabajo infantil, con el impacto en las comunidades locales, la corrupción y las políticas públicas. El segundo está relacionado con los materiales utilizados, el consumo de energía, con las emisiones y residuos, con los productos y servicios, con la evaluación y las medidas correctivas en el campo de los derechos humanos, y el impacto en las comunidades locales.

En la Figura 4, las empresas en cada clúster se han coloreado respectivamente y se ha dibujado un polígono conteniendo todas las empresas. Cuatro clúster han sido descritos y representados en el mapa euclídeo de esta figura.

De acuerdo a nuestros resultados, la información más relevante recogida con el modelo sería:

- El clúster 3, representado en color morado, está formado por las 91 empresas (24,14%) más sostenibles. Cabe destacar que el 100% de las empresas del clúster 3 publican los índices EC1, LA1, LA5, LA7 y LA13. Más del 90% publican los índices EC2, EC3, EC4, EC6, EC7, EC8, EN3, EN4, EN8, EN16, EN22, EN26, EN28, LA2, LA4, LA8, LA10, LA14, HR1, HR2, HR4, HR5, HR6, HR7, HR11, SO1, SO2, SO3, SO4, SO5, SO8 y PR9. Entre el 80% y el 89% publican los índices EN1, EN17, EN21, EN23, LA15, HR3, HR10, SO9, SO10, PR1, PR3 y PR6. Entre el 70% y el 79% publican los índices EN2, EN11, EN12, EN20. Menos del 60% publican los índices EN19 y EN27.
- El clúster 4, representado en color verde, contiene 149 empresas (39,52%) con un nivel de sostenibilidad intermedio-alto. Observamos que más del 90% de las empresas publican los índices EC1, EC3, EC6, EC8, EN3, EN4, EN8, EN16, EN22, EN26, LA1, LA2, LA4, LA7, LA8, LA10, LA13, SO2, SO3, PR3. Entre el 80% y el 89% publican los índices EC2, EC4, EC7, EN1, EN2, EN11, EN12, EN17, EN19, EN20, EN21, EN28, LA5, LA14, HR2, HR4, HR5, HR6, HR7, SO1, SO4, SO5, SO8, PR1, PR6, PR9. Entre el 50% y el 75% publican los índices EN23, EN27, HR1, HR3. Menos del 25% publican los índices LA15, HR10, HR11, SO9, SO10. No obstante, debe tenerse en cuenta que estos índices se añadieron en la guía G3.1.

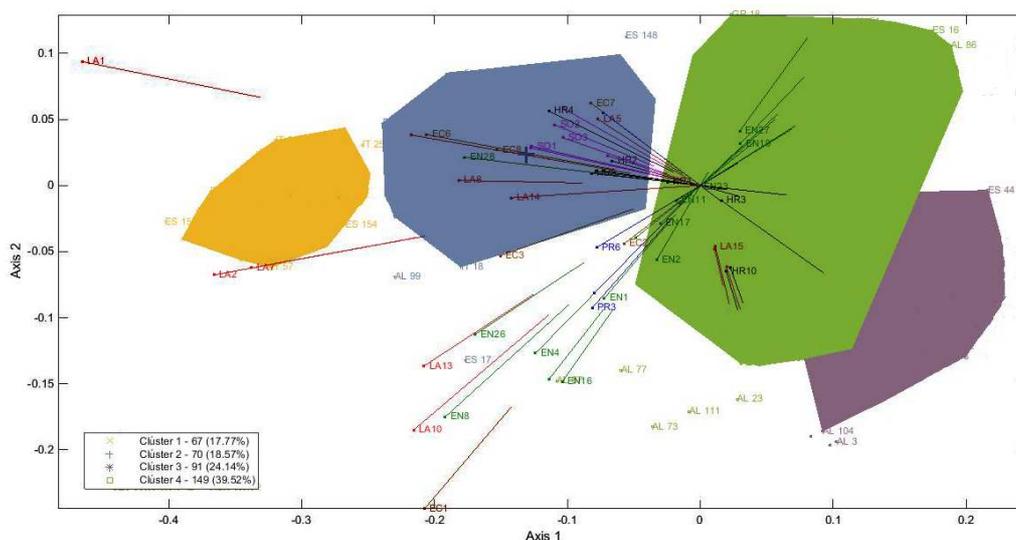


Figura 4. Biplot logístico de dos dimensiones con clúster, basado en los datos de sostenibilidad de empresas de Alemania y países PIIGS. Año 2011. Elaborado por Belén Vaquero Romero.

- El clúster 2, representado en color azul, contiene 70 empresas (18,57%) con un nivel intermedio-bajo de sostenibilidad. De ellas, más de un 90% publican los índices EC1, LA1, LA2, LA7 y entre el 80% y el 89% publican los índices EN3, EN8, EN16, EN22, LA10, LA13. Entre el 70% y el 79% publican los índices EN4, EN26, LA4 y entre el 50% y el 69% publican los índices EC3, EC4, EN1, EN2, EN28, LA8, LA14, HR4, PR1, PR3. Menos del 50% de las empresas publican el resto de los índices GRI.
- El clúster 1, representado en color amarillo, contiene las 67 empresas (17,77%) con menor nivel de sostenibilidad. Únicamente los índices EC1 y LA1 están publicados por más del 70% de las empresas. Entre el 50% y el 65% publican los índices EN3 y LA10. El resto de índices están publicados por un porcentaje de empresas significativamente menor.

A continuación, estudiaremos las diferencias en sostenibilidad entre Alemania y los llamados **países PIIGS**, en el **año 2012**. El ajuste de bondad global como porcentaje de clasificación correcta en el biplot es 84,88%, por lo que con una solución de dos dimensiones será suficiente para explorar las características principales de los datos.

La solución biplot obtenida se muestra en la Figura 5. Todas las variables están significativamente relacionadas con la ordenación observada de las empresas ya que todas tienen un p-valor menor que 0,05. Las variables más relevantes para la interpretación de la ordenación de las empresas en el diagrama de dispersión son aquellas con mayor R^2 . En nuestro caso, estas variables son: HR10 (0,83), HR11 (0,86), SO9 (0,93) y SO10 (0,93):

El gradiente asociado al eje 1 está principalmente relacionado con las variables EC1, EC2, EC3, EC6, EC7, (Económicas), EN8, EN20, EN21, EN22, EN23, EN28 (Medioambientales), LA4, LA5, LA7, LA8, LA14, (Laborales), HR1, HR2, HR4, HR5, HR6, HR7, (Recursos Humanos), SO3, SO4, SO5, SO8, (Sociales), PR6, PR9, (Responsabilidad de Productos): El gradiente asociado al eje 2 está relacionado con las variables EN1, EN2, (Medioambientales), LA1, LA15, (Laborales), HR10, HR11, (Recursos Humanos), SO9, SO10, (Sociales).

Hemos encontrado dos gradientes principales. El primero está relacionado con el desempeño económico y la presencia en el mercado, con el consumo de energía, con las emisiones, vertidos y residuos, con el cumplimiento normativo, con el empleo, con las relaciones empresa-trabajadores, con la salud y la seguridad laboral, con las prácticas de inversión y abastecimiento, con la no discriminación, con la libertad de asociación, con el trabajo infantil y los trabajos forzados, con la corrupción y las políticas públicas, con las comunicaciones de marketing y la privacidad del cliente. El segundo está relacionado con los materiales utilizados, con el empleo, con la evaluación y las medidas correctivas en el campo de los derechos humanos, y con el impacto en las comunidades locales.

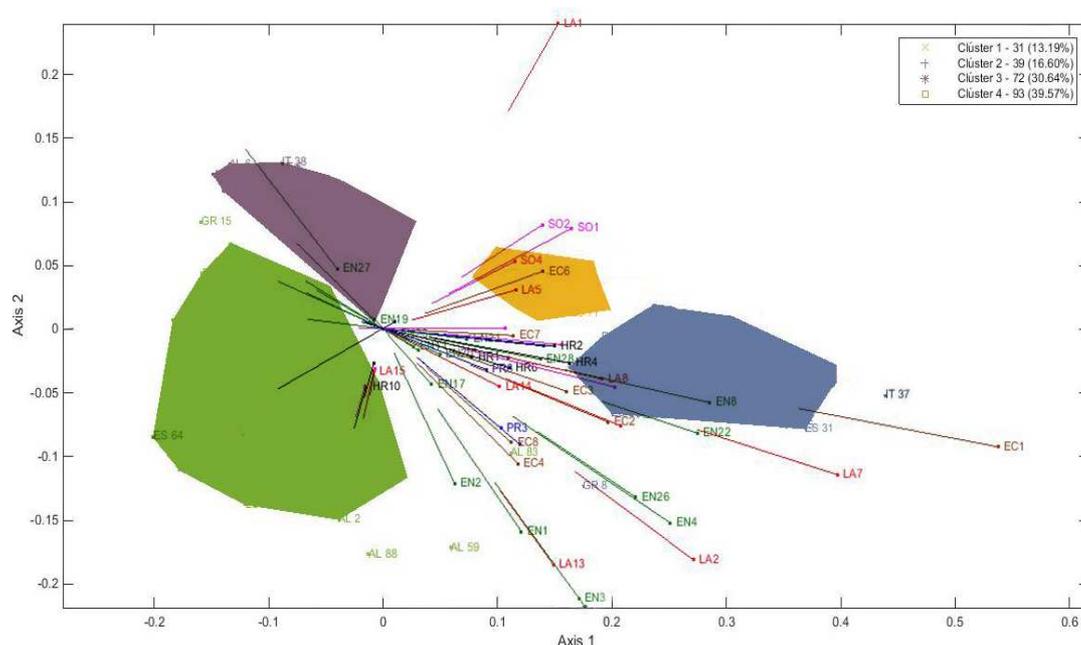


Figura 5. Biplot Logístico Externo de dos dimensiones con clúster, basado en los datos de sostenibilidad de empresas de Alemania y los llamados países PIIGS. Año 2012. Elaborado por Belén Vaquero Romero.

- El clúster 4, representado en color verde, está formado por las 93 empresas (39,57%) con mayor sostenibilidad. Cabe destacar que el 100% de las empresas del clúster 4 publican el índice LA1 y más del 90% publican los índices EC1, EC2, EC3, EC6, EC7, EN3, EN4, EN8, EN16, EN22, EN26, EN28, LA2, LA4, LA5, LA7, LA8, LA10, LA13, HR5, HR7, SO1, SO2, SO3, SO4, SO5. Entre el 80% y el 89% publican los índices EC4, EC8, EN1, LA14, HR1, HR2, HR4, HR6, HR11, SO8, SO9, SO10, PR1, PR3, PR6 y PR9. Entre el 70% y el 79% publican los índices EN2, EN11, EN12, EN17, EN20, EN21, EN23, LA15, HR3 y HR10. Entre el 40% y el 60% publican los índices EN19 y EN27.
- El clúster 3, representado en color morado, contiene a las 72 empresas (30,64%) que presentan un nivel de sostenibilidad medio-alto. El 100% de las empresas que en él se incluyen publican el índice LA13 y más del 90% publican los índices EC1, EC2, EC3, EC8, EN1, EN2, EN3, EN4, EN8, EN16, EN22, EN26, EN28, LA1, LA2, LA4, LA7, LA8, LA10, HR2, HR4, SO2, SO3, SO4, SO5, PR1, PR3. Entre el 80% y el 89% publican los índices EC4, EC6, EC7, EN17, LA5, LA14, HR1, HR5, HR6, HR7, SO1, SO8, PR6, PR9. Entre el 70% y el 79% publican los índices EN11, EN12, EN21, EN23. Entre el 50% y el 69% publican los índices EN19, EN20, EN27, HR3. Menos del 5% publican los índices LA15, HR10, HR11, SO9, SO10. No obstante, debe tenerse en cuenta que estos índices se añadieron en la guía G3.1.
- El clúster 1, representado en color amarillo, contiene a las 31 empresas (13,19%) con nivel de sostenibilidad medio-bajo. De ellas, más de un 90% publican los índices EC1, EN8, EN16, EN22, LA7 y LA13, y entre el 70% y el 89% publican los índices EC2, EN1, EN3, EN4, EN26, LA1, LA2, LA8, LA10 y PR1. Entre el 50% y el 69% publican los índices EC3, EC4, EC8, EN2, EN17, LA4, LA5, LA14, HR2, HR4, HR6, SO1, SO5 y PR3. Entre el 40% y el 49% publican los índices EC6, EC7, EN20, EN21, HR7, SO2, SO3, PR6 y PR9, y menos del 39% de las empresas publican el resto de los índices.
- El clúster 2, representado en color azul, contiene a las 39 empresas (16,59%) con menor nivel de sostenibilidad. De ellas, más de un 80% publican los índices EC1 y LA1. Entre el 60% y el 79% publican los índices EN3, EN16, LA2, LA7, LA10, LA13. Entre el 40% y el 59% publican los índices EN1, EN2, EN4, EN8, EN22, EN26, LA4. El resto de índices son publicados por menos del 39% de las empresas.

En base a nuestros resultados, tanto en el año 2011 como en el 2012, las empresas españolas presentan el porcentaje más elevado de publicaciones en los clústeres de máxima sostenibilidad, seguidas de las empresas

alemanas. No obstante, si nos fijamos en el porcentaje de las empresas más sostenibles dentro de cada país, Portugal presenta una clara ventaja a pesar del escaso número de informes que publica. En ambos años, casi la mitad de sus empresas se encuentran dentro del clúster de máxima sostenibilidad.

Además, nuestro estudio nos destaca que las empresas más sostenibles son, en su mayoría, de gran tamaño, privadas y pertenecientes al sector servicios. Dentro de este grupo de empresas hay una clara tendencia a la utilización de las directrices de la guía G3.1, haciendo uso de la comprobación GRI y del nivel de aplicación A+ en la publicación de sus informes. (Ver Anexos Tablas 11, 12, 13 y 14),

Respecto al clúster con menor sostenibilidad, en ambos años estaría formado por una mayoría de empresas de España, que también presenta unos niveles muy altos de inclusión en el mencionado clúster dentro del propio país. Vemos por tanto que, a pesar del gran número de informes que las empresas españolas publican cada año, su modelo debe aún progresar hacia niveles más sostenibles.

En las Tablas 15, 16, 17 y 18 de los Anexos, se presenta la clasificación de las empresas de Grecia e Italia con mayor nivel de sostenibilidad para los años 2011 y 2012. Además se presenta en la Tabla 19 de los Anexos la clasificación global de las empresas con mayor nivel de sostenibilidad para los años 2011 y 2012.

4. DISCUSIÓN DE LOS RESULTADOS

Nuestro estudio se ha centrado, precisamente, en conocer la sostenibilidad económica, ambiental y social que presentan las empresas europeas, a través del índice GRI. El número de empresas que publican informes de sostenibilidad de acuerdo a las directrices GRI ha experimentado un crecimiento notable desde la aparición de las primeras guías. Este crecimiento es especialmente significativo en Europa, ya que los clientes y empleados son considerados como una parte muy relevante dentro de los grupos de interés que se pueden ver afectados por las actividades de las empresas (Marimon, Alonso-Almeida, Del Pilar Rodríguez y Cortez, 2012). De manera general, se ha visto que los indicadores más publicados por las empresas europeas han sido los relacionados con la parte *económica* EC1 (Valor económico directo generado y distribuido), la parte *medio ambiental* EN3 (Consumo directo de energía) y EN16 (Emisiones totales de gases de efecto invernadero) y la parte *laboral* LA1 (Desglose del colectivo de trabajadores por tipo de empleo, por contrato y por región), LA7 (Tasas de absentismo, enfermedades profesionales, días perdidos y número de víctimas mortales relacionadas con el trabajo por región), y LA13 (Composición de los órganos de gobierno corporativo y plantilla, desglosado por sexo, grupo de edad, pertenencia a minorías y otros indicadores de diversidad): En un estudio previo realizado por Anghel-Ilcu (2012), en el que se comparaban los hábitos de sostenibilidad entre empresas anglosajonas, representadas por Gran Bretaña, y empresas de Europa continental, representadas por Francia, destacan los altos porcentajes de publicación de los mismos indicadores obtenidos en nuestra investigación. La autora afirma que a pesar de la variedad de indicadores económicos, sociales y ambientales propuestos por GRI, los que finalmente son publicados por las empresas suelen concentrarse en un número limitado de temas.

En general, se detecta escasez en la publicación de indicadores sociales por las empresas debido a la falta de regulación en éste área, lo que podría ser la razón por la que las compañías tienden menos a la publicación de este tipo de indicadores en comparación con los ambientales o los laborales. Se hace por tanto evidente la importancia de implementar una regulación en el área de la publicación de indicadores sociales no solo relacionados con los trabajadores, sino también con los clientes y proveedores (Bonsón, y Bednárová, 2014).

Una parte fundamental de nuestro estudio ha sido el análisis detallado por tipo de país, donde hemos distinguido entre aquellos países con mayor PIB per cápita de Europa y aquellos más afectados por la crisis financiera actual, conocidos como PIIGS (Siemiatkowski y Jankowska, 2013). Por un lado, como muestra representativa de los países con mayor poder adquisitivo de Europa hemos elegido a Alemania, Países Bajos, Suecia y Suiza. Una vez procesados los datos, observamos que Alemania es, además del país con mayor número de publicaciones, el país con mayores niveles de sostenibilidad. Por el contrario, Suecia resulta ser el país menos comprometido en materia de sostenibilidad, a pesar de los numerosos informes que publica cada año.

De acuerdo a Gamerschlag Möller, y Verbeeten (2011), el hecho de que las empresas alemanas voluntariamente exhiban sus logros y actividades en los informes de sostenibilidad está asociado a una mayor visibilidad de las compañías, a una estructura de propiedad más dispersa entre los accionistas y a la influencia de los grupos de interés norteamericanos. El alto nivel de sostenibilidad registrado por Alemania podría estar relacionado con estos factores.

En el caso de Suecia, nuestros resultados coinciden con la opinión de autores como Hąbek y Wolniak (2013), los cuales indican que el gran número de memorias de sostenibilidad GRI que el país publica cada año se debe a la normativa que el gobierno introdujo en el año 2008, a través de la cual se hacía obligatoria la publicación de informes de sostenibilidad por parte de todas las empresas propiedad del estado. Esto ayudó a incrementar la conciencia de las compañías a este respecto y mejoró los procesos internos relacionados con la preparación de los informes pero no introdujo cambios en las prácticas por parte de las empresas, por lo que los resultados de los informes no presentan el nivel de sostenibilidad que cabría esperar.

Coincidiendo con nuestros resultados, Marimon et al. (2012), indican que a nivel general las directrices GRI para la publicación de informes de sostenibilidad son adoptadas por una mayoría de empresas de gran tamaño, ya que las pequeñas y medianas encuentran menos incentivos en un proceso que les puede resultar caro y complejo.

En la parte de nuestro estudio dedicada a los países más afectados por la crisis económica, realizamos una comparación entre dichos países y Alemania, con el objetivo de conocer si el poder adquisitivo de los países influye en su compromiso con la sostenibilidad. En los dos años analizados, Portugal figura como el país más sostenible a pesar del limitado número de informes publicados, seguido por Italia, España y Grecia, quedando Alemania como el país con menos empresas sostenibles. De acuerdo a (Castelo Branco y col., 2014), un factor que ha contribuido al descenso de informes de sostenibilidad publicados por Portugal ha sido la deuda pública que afecta de gran manera a las empresas propiedad del estado ya que, para implicarse en actividades que promuevan el desarrollo sostenible y la publicación de informes de sostenibilidad, es necesario disponer de recursos económicos dentro de la compañía. Esto ha producido una disminución tanto en el número de informes, como en la verificación externa de las memorias.

Respecto a la calidad de las memorias presentadas por los países en crisis, Gallén-Ortiz y Giner-Inchausti (2013) detallan que en el periodo comprendido entre el año 2007 y el 2011, los países mejor posicionados serían Portugal e Italia, detectando por el contrario una caída en la calidad de los informes de España. Estas conclusiones coincidirían con nuestros resultados. En esta misma línea se encuentran los resultados de Vicente Galindo y col. (2015), obtenidos a partir del estudio de informes de sostenibilidad de empresas portuguesas y españolas a lo largo del año 2011. De acuerdo a las autoras, si bien el número de publicaciones de España es significativamente mayor, Portugal destaca por la calidad de sus informes, resultando ser notablemente más sostenible que España.

De acuerdo a Giannarakis y Theotokas (2011), si bien la crisis financiera ha puesto en duda si los informes de sostenibilidad constituyen una amenaza para la supervivencia de la empresa debido a sus altos costes, la conclusión es que la crisis actual ofrece la posibilidad a las empresas de redirigir esta situación, viendo a los informes de sostenibilidad no como una amenaza, sino como una oportunidad. Los beneficios que pueden surgir con la implementación de medidas e iniciativas para la publicación de informes de sostenibilidad pueden ayudar a las empresas a diferenciar sus bienes y servicios y reestablecer la confianza entre las compañías y sus grupos de interés.

Para consultar más detalles ver Vaquero-Romero (2015):

5. CONCLUSIONES

En base a la bibliografía consultada, y a los resultados obtenidos en nuestro estudio, concluimos que:

- La elaboración de informes de sostenibilidad de acuerdo a las directrices marcadas por la Global Reporting Initiative (GRI) ha experimentado un gran aumento desde sus comienzos en el año 1999 hasta la actualidad. La importancia que otorga Europa a los grupos de interés de cada empresa ha hecho que el número de publicaciones en esta región destaque de manera tan significativa.
- De los indicadores GRI analizados, se observa una tendencia clara hacia la publicación de aquellos relacionados con la parte económica, ambiental y laboral. La escasez en la publicación de indicadores sociales puede deberse a la falta de regulación en éste área.
- La publicación de informes de sostenibilidad resulta de gran importancia tanto para los países con alto poder adquisitivo, como para aquellos que se han visto más afectados por la crisis económica actual. Las empresas más sostenibles de todo el estudio son en su mayoría de gran tamaño y privadas. Las pequeñas y medianas empresas, si bien están presentes, son minoritarias al encontrar menos incentivos en un proceso que resulta caro y complejo.
- Portugal resulta ser el país más sostenible de entre todos los analizados, si bien el número de informes de sostenibilidad publicados es significativamente bajo. Esta situación estaría

estrechamente relacionada con la falta de recursos económicos para destinar a este tipo de actividades en un momento de crisis como el actual.

- Destaca el caso de países como Suecia que, a pesar de su solvencia económica y del alto número de informes que registra, presenta una calidad muy baja en sus declaraciones de sostenibilidad.

El análisis llevado a cabo, en el que se ha estudiado de manera general la situación global europea y, con más detenimiento, las diferencias entre países afectados en mayor o menor medida por la crisis financiera actual, nos lleva a concluir que Europa muestra un gran compromiso en materia de publicación de informes de sostenibilidad, y que la crisis económica no ha afectado a la calidad de las publicaciones, mostrando las empresas una responsabilidad a este respecto incluso mayor al encontrarse en épocas de dificultad.

REFERENCIAS

- [1] Anghel-Ilcu, E.R. (2012): European evidence regarding assurance statement practices in social and environmental reporting. 7th International Conference Accounting and Management Information Systems AMIS 2012, Conferencia llevada a cabo en Bucarest.
- [2] Datos.bancomundial.org. (2016): PIB per cápita (US\$ a precios actuales): Recuperado de: <http://datos.bancomundial.org/indicador/NY.GDP.PCAP.CD>
- [3] Bonsón, E. y M. Bednárová (2015): CSR reporting practices of Eurozone companies, *Revista de Contabilidad*, 18 (2), p. 182-193.
- [4] Castelo Branco, M., C. Delgado, S. Ferreira Gomes y T.C. Pereira Eugénio (2014): Factors influencing the assurance of sustainability reports in the context of the *economic* crisis in Portugal, *Managerial Auditing Journal*, 29(3), 237-252.
- [5] Demey, J., J. L. Vicente-Villardón, M. P. Galindo y A. Zambrano (2008): Identifying molecular markers associated with classification of genotypes using external logistic biplots, *Bioinformatics*, 24(24), 28-32.
- [6] Fundación Etnor-EEI Centro Europeo de Empresas Innovadoras (2004): Guía de responsabilidad social de la empresa. Recuperado de: <http://goo.gl/ENkdGX>
- [7] Gallén Ortiz, M.L. y B. Giner Inchausti (2013): Los efectos de la crisis económica en las memorias de sostenibilidad GRI en Europa, *Prisma Social*, 10, 31-63.
- [8] Gamerschlag, R., K. Möller y F. Verbeeten (2011): Determinants of voluntary CSR disclosure: empirical evidence from Germany, *Review of Managerial Science*, 2011 (5), 233-262.
- [9] Giannarakis, G. e I. Theotokas (2011): The Effect of Financial Crisis in Corporate Social Responsibility Performance, *International Journal of Marketing Studies*, 3(1):
- [10] Global Reporting Initiative (2011): Guía para la elaboración de Memorias de Sostenibilidad. Versión 3.1. Amsterdam, The Netherlands. Recuperado de: <https://goo.gl/nRLMKJ>.
- [11] Global Reporting Initiative. (2012): La elaboración de memorias de sostenibilidad de GRI: ¿Vale la pena el viaje?. Recuperado de: <https://www.globalreporting.org/>
- [12] Global Reporting Initiative. (2006): Guía para la elaboración de Memorias de Sostenibilidad (2000-2006): Recuperado de: <https://www.globalreporting.org/>
- [13] Hąbek, P. y R. Wolniak (2013): European Union regulatory requirements relating to Sustainability Reporting. The case of Sweden, *Scientific Journals of the Maritime University of Szczecin*, 34(106), 40-47.
- [14] Marimon, F., M. M. Alonso-Almeida, M. Del Pilar Rodríguez y K.A. Cortez Alejandro (2012): The worldwide diffusion of the global reporting initiative: what is the point?, *Journal of Cleaner Production*, 33, 132-144.
- [15] Moneva Abadía, J.M. (2007): El marco de la información sobre responsabilidad social de las organizaciones, *Ekonomiaz: Revista vasca de economía*, 65, 284-317.
- [16] Parlamento Europeo (2006): Nomenclatura estadística de actividades económicas NACE. Revisión 2. Recuperado de: <http://goo.gl/9zOFET>.
- [17] Siemiakowski, P. y E. Jankowska (2013): Financial dependence of the PIIGS countries, *Journal of World Economic Research*, 2(5), 89-94.
- [18] UNCSO. (2012): Rio+20-United Nations Conference on Sustainable Development. Recuperado de: <http://www.uncsd2012.org/>
- [19] United Nations Statistics Division-Standard Country and Area Codes Classifications (2013): Recuperado de: <http://bit.ly/1kEneGK>.

- [20] Valor Martínez, C. y M. De la Cuesta González (2003): Responsabilidad Social de la Empresa: Concepto, Medición y Desarrollo en España, *Boletín Económico ICE*, 2755, 7-20.
- [21] Vaquero-Romero, B. (2015): Sostenibilidad económica, ambiental y social en las empresas europeas. Análisis multivariante del índice GRI. TFG. Facultad de Ciencias. USAL. España.
- [22] Vicente Galindo, P., E. Vaz y T. de Noronha (2015): How Corporations Deal with Reporting Sustainability: Assessment Using the Multicriteria Logistic Biplot Approach, *Systems*, 3(1), 6-26.
- [23] Vicente-Villardón, J.L. (2015): MULTBILOT: A package for Multivariate Analysis using Biplots. Departamento de Estadística. Universidad de Salamanca. Recuperado de: <http://biplot.usal.es/ClassicalBiplot/index.html>
- [24] Waas, T., J. Hugé, A. Verbruggen y T. Wright (2011): Sustainable Development: A Bird's Eye View, *Sustainability*, 3(10), 1637-1661.
- [25] White, A.L. (1999): Sustainability and the Accountable Corporation, *Environment: Science and Policy for Sustainable Development*, 41(8), 30-43.
- [1] Wyplosz, C. (2010): The Eurozone in the Current Crisis. ADBI Working Paper 207. Tokyo: Asian Development Bank Institute.

6. ANEXOS

Tabla 3. Clasificación de las empresas de Alemania con mayor nivel de sostenibilidad. Año 2011.

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
COCA-COLA DEUTSCHLAND	Grande	Privada	Industrial	Fabricación de bebidas	100,00%
FRAPORT AG	Grande	Privada	Servicios	Transporte aéreo	100,00%
MUNICH AIRPORT	Grande	Privada	Servicios	Transporte aéreo	100,00%
REWE GROUP	Grande	Privada	Servicios	Comercio al por menor	100,00%
SANLUCAR FRUITS	Grande	Privada	Primario	Agricultura	100,00%
TYCZKA UNTERNEHMENSGRUPPE	Grande	Privada	Agua y Energía	Suministro de energía	100,00%
AXEL SPRINGER	Grande	Privada	Servicios	Actividades editoriales	98,18%
SOLAR WORLD	Grande	Privada	Agua y Energía	Suministro de energía	98,18%
ENTEKA	Grande	Filial	Agua y Energía	Suministro de energía	96,36%
HEAG SÜDHESSISCHE ENERGIE AG (HSE)	Grande	Propiedad del Estado	Agua y Energía	Suministro de energía	96,36%

Tabla 4. Clasificación de las empresas de Alemania con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
FRAPORT AG	Grande	Privada	Servicios	Transporte aéreo	100,00%
MERCK GERMANY	Mediana	Privada	Industrial	Fabricación de productos farmacéuticos	100,00%
MUNICH AIRPORT	Grande	Privada	Servicios	Transporte aéreo	100,00%
REWE GROUP	Grande	Privada	Servicios	Comercio al por menor	100,00%
SANLUCAR FRUITS	Grande	Privada	Primario	Agricultura	100,00%
TCHIBO	Grande	Privada	Servicios	Comercio al por menor	100,00%
BMW GROUP	Grande	Privada	Industrial	Fabricación de vehículos de motor	98,18%
BAYER AG	Mediana	Privada	Servicios	Actividades profesionales, científicas y técnicas	96,36%
DEUTSCHE BAHN	Grande	Propiedad del Estado	Servicios	Transporte terrestre	96,36%
ENTEKA	Grande	Filial	Agua y Energía	Suministro de energía	96,36%

Tabla 5. Clasificación de las empresas de Suecia con mayor nivel de sostenibilidad. Año 2011

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
HOLMEN	Grande	Privada	Industrial	Industria del papel	100,00%
BYGGMAX	Grande	Privada	Servicios	Comercio al por menor	90,90%
ERICSSON	Grande	Privada	Servicios	Telecomunicaciones	90,90%
SAS GROUP AB (SWEDEN)	Grande	Privada	Servicios	Transporte aéreo	90,90%
SCA - SVENSKA CELLULOSA AKTIEBOLAGET	Grande	Privada	Industrial	Industria del papel	90,90%
SKF GROUP	Grande	Privada	Industrial	Industrias manufactureras	90,90%
SYSTEMBOLAGET	Grande	Privada	Servicios	Comercio al por menor	90,90%
AAK	Grande	Privada	Industrial	Industria de la alimentación	89,09%
ATRIUM LJUNGBERG	Grande	Privada	Servicios	Actividades inmobiliarias	89,09%
NORDEA BANK	Grande	Privada	Servicios	Actividades financieras	89,09%

Tabla 6. Clasificación de las empresas de Suecia con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
HOLMEN	Grande	Privada	Industrial	Industria del papel	100,00%
BYGGMAX	Grande	Privada	Servicios	Comercio al por menor	90,90%
ERICSSON	Grande	Privada	Servicios	Telecomunicaciones	90,90%
FENIX OUTDOOR	Mediana	Privada	Servicios	Comercio al por menor	90,90%
SCA - SVENSKA CELLULOSA AKTIEBOLAGET	Grande	Privada	Industrial	Industria del papel	90,90%
SKF GROUP	Grande	Privada	Industrial	Industrias manufactureras	90,90%
SYSTEMBOLAGET	Grande	Propiedad del Estado	Servicios	Comercio al por menor	90,90%
AAK	Grande	Privada	Industrial	Industria de la alimentación	89,09%
AXFOOD	Grande	Privada	Servicios	Comercio al por menor	89,09%
SAS GROUP AB (SWEDEN)	Grande	Privada	Servicios	Transporte aéreo	89,09%

Tabla 7. Clasificación de las empresas de los Países Bajos con mayor nivel de sostenibilidad. Años 2011

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
ENEXIS	Grande	Propiedad del Estado	Agua y Energía	Suministro de energía	98,18%
RICOH	Grande	Privada	Servicios	Actividades relacionadas con la informática	96,36%
ACHMEA	Grande	Filial	Servicios	Actividades de seguros	94,54%
TNO	Grande	Sin ánimo de lucro	Otros Servicios	Otros servicios	94,54%
TRANSPARABILITY B.V.	Pequeña	Privada	Servicios	Actividades profesionales, científicas y técnicas	90,90%
ASN BANK	Pequeña	Filial	Servicios	Actividades financieras	89,09%
MN	Mediana	Privada	Servicios	Actividades financieras	89,09%
PHILIPS	Mediana	Privada	Industrial	Industrias manufactureras	89,09%
POSTNL	Mediana	Privada	Servicios	Actividades postales y de correos	89,09%
REDEVCO	Grande	Privada	Servicios	Actividades inmobiliarias	89,09%

Tabla 8. Clasificación de las empresas de los Países Bajos con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
CROWN VAN GELDER	Grande	Privada	Industrial	Industria del papel	100,00%
POSTNL	Mediana	Privada	Servicios	Actividades postales y de correos	98,18%
WAVIN	Mediana	Privada	Industrial	Fabricación de productos de caucho y plásticos	98,18%
KPN	Grande	Privada	Servicios	Telecomunicaciones	96,36%
ACHMEA	Grande	Filial	Servicios	Actividades de Seguros	94,54%
ASML	Grande	Privada	Industrial	Fabricación de productos informáticos	92,72%
SCHIPHOL	Grande	Propiedad del Estado	Servicios	Transporte aéreo	92,72%
VAN GANSEWINKEL GROUP	Mediana	Privada	Agua y Energía	Tratamiento de aguas residuales	90,90%
ASN BANK	Pequeña	Filial	Servicios	Actividades financieras	89,09%
MN	Mediana	Privada	Servicios	Actividades financieras	89,09%

Tabla 9. Clasificación de las empresas de Suiza con mayor nivel de sostenibilidad. Año 2011

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
MCI	Grande	Privada	Servicios	Actividades profesionales, científicas y técnicas	98,18%
ABB ASEA BROWN BOVERI LTD	Mediana	Privada	Industrial	Fabricación de material eléctrico	96,36%
NESTLÉ	Mediana	Privada	Industrial	Industria de la alimentación	94,54%
SYNGENTA	Mediana	Privada	Industrial	Industria química	92,72%
CLARIANT	Grande	Privada	Industrial	Industria química	90,90%
AXPO HOLDING	Grande	Privada	Agua y Energía	Suministro de energía	89,09%
GRAUBÜNDNER KANTONALBANK (GKB)	Grande	Privada	Servicios	Actividades financieras	89,09%
GEBERTIT	Mediana	Privada	Industrial	Industrias manufactureras	87,27%
GEORG FISCHER	Grande	Privada	Industrial	Industrias manufactureras	85,45%
NOVARTIS	Mediana	Privada	Industrial	Fabricación de productos farmacéuticos	85,45%

Tabla 10. Clasificación de las empresas de Suiza con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
ABB ASEA BROWN BOVERI LTD	Mediana	Privada	Industrial	Fabricación de material eléctrico	96,36%
NESTLÉ	Mediana	Privada	Industrial	Industria de la alimentación	96,36%
SYNGENTA	Mediana	Privada	Industrial	Industria química	92,72%
CLARIANT	Grande	Privada	Industrial	Industria química	90,90%
AXPO HOLDING	Grande	Privada	Agua y Energía	Suministro de energía	89,09%
GLENCORE XSTRATA	Mediana	Privada	Primario	Extracción de minerales metálicos	89,09%
WORLD VISION INTERNATIONAL	Grande	Organización sin ánimo de lucro	Otros Servicios	Actividades asociativas	89,09%
GEBERTIT	Mediana	Privada	Industrial	Industrias manufactureras	87,27%
GRAUBÜNDNER KANTONALBANK (GKB)	Grande	Privada	Servicios	Actividades financieras	87,27%
MIGROS	Grande	Cooperativa	Servicios	Comercio al por menor	87,27%

Tabla 11. Clasificación de las empresas de España con mayor nivel de sostenibilidad. Año 2011

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
ABEINSA	Grande	Privada	Agua y Energía	Ingeniería y construcción	100,00%
ABERTIS	Grande	Privada	Otros	Otros servicios	100,00%
BANCO SABADELL	Grande	Privada	Servicios	Actividades financieras	100,00%
ENDESA	Grande	Privada	Otros	Suministro de energía	100,00%
GRUPO ACS	Grande	Privada	Construcción	Construcción	100,00%
GRUPO UNIÓN FENOSA GAS	Mediana	Privada	Agua y Energía	Suministro de energía	100,00%
TELVENT	Grande	Privada	Servicios	Telecomunicaciones	100,00%
ABENGOA SOLAR	Grande	Privada	Agua y Energía	Suministro de energía	98,18%
ACCENTURE SPAIN	Grande	Privada	Servicios	Actividades de consultoría	98,18%
AQUAGEST REGIÓN DE MURCIA	Grande	Cooperativa	Agua y Energía	Suministro de agua	98,18%

Tabla 12. Clasificación de las empresas de España con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
ABEINSA	Grande	Privada	Agua y Energía	Ingeniería y construcción	100,00%
AGBAR	Grande	Privada	Agua y Energía	Suministro de agua	100,00%
AQUALIA	Grande	Privada	Agua y Energía	Suministro de agua	100,00%
BANCO SABADELL	Grande	Privada	Servicios	Actividades financieras	100,00%
BANKINTER	Grande	Privada	Servicios	Actividades financieras	100,00%
ENDESA	Grande	Privada	Agua y Energía	Suministro de energía	100,00%
RENFE	Grande	Propiedad pública	Servicios	Transporte terrestre	100,00%
ABERTIS	Grande	Privada	Otros	Otros servicios	98,18%
HEINEKEN SPAIN	Grande	Privada	Industrial	Fabricación de bebidas	98,18%
INDRA	Grande	Privada	Servicios	Actividades informáticas	98,18%

Tabla 13. Clasificación de las empresas de Portugal con mayor nivel de sostenibilidad. Año 2011

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
ÁGUAS DO DOURO E PAIVA	Grande	Propiedad del Estado	Agua y Energía	Suministro de agua	100,00%
ANA	Grande	Privada	Servicios	Transporte aéreo	100,00%
AXA SEGUROS PORTUGAL	Grande	Privada	Servicios	Actividades de seguros	100,00%
BANCO ESPIRITO SANTO	Grande	Privada	Servicios	Actividades financieras	100,00%
BRISA	Grande	Privada	Servicios	Transporte terrestre	100,00%
GRUPO PORTUCEL	Grande	Privada	Industrial	Industria del papel	100,00%
PORTO DE SINES	Mediana	Privada	Servicios	Actividades relacionadas con el transporte	98,18%
ÁGUAS DE PORTUGAL	Grande	Propiedad del Estado	Agua y Energía	Suministro de agua	96,36%
GRUPO AUCHAN	Grande	Privada	Servicios	Comercio al por menor	96,36%
LIPOR	Grande	Propiedad Pública	Agua y Energía	Gestión de residuos	96,36%

Tabla 14. Clasificación de las empresas de Portugal con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
ANA	Grande	Privada	Servicios	Transporte aéreo	100,00%
ÁGUAS DO DOURO E PAIVA	Grande	Propiedad del Estado	Agua y Energía	Suministro de agua	96,36%
LIPOR	Grande	Propiedad Pública	Agua y Energía	Gestión de residuos	96,36%
EDP (Energías de Portugal)	Grande	Privada	Agua y Energía	Suministro de energía	92,73%
BANCO ESPIRITO SANTO	Grande	Privada	Servicios	Actividades financieras	90,91%
BRISA	Grande	Privada	Servicios	Transporte terrestre	90,91%
SONAECOM	Grande	Privada	Servicios	Actividades informáticas	90,91%
NESTLÉ PORTUGAL	Grande	Filial	Industrial	Industria de la alimentación	90,91%
SONAE SIERRA	Grande	Privada	Servicios	Actividades inmobiliarias	89,09%
PORTUGAL TELECOM	Grande	Privada	Servicios	Telecomunicaciones	85,45%

Tabla 15. Clasificación de las empresas de Grecia con mayor nivel de sostenibilidad. Año 2011

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
HELLENIC CABLES	Mediana	Privada	Industrial	Industrias manufactureras	100,00%
CORINTH PIPEWORKS	Grande	Privada	Industrial	Industrias manufactureras	96,36%
ELVAL	Grande	Privada	Industrial	Industria metalúrgica	96,36%
HALCOR	Grande	Privada	Industrial	Industria metalúrgica	96,36%
HELLENIC PETROLEUM	Grande	Privada	Agua y Energía	Suministro de energía	89,09%
COSMOTE MOBILE TELECOMMUNICATIONS S.A.	Grande	Filial	Servicios	Telecomunicaciones	85,45%
TITAN CEMENT	Grande	Privada	Construcción	Materiales de construcción	85,45%
FRIGOGLASS	Mediana	Privada	Industrial	Fabricación de bebidas	85,45%
PIRAEUS BANK	Mediana	Privada	Servicios	Actividades financieras	81,82%
GEFYRA	Pequeña	Privada	Construcción	Ingeniería civil	80,00%

Tabla 16. Clasificación de las empresas de Grecia con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
HELLENIC CABLES	Mediana	Privada	Industrial	Industrias manufactureras	100,00%
ELVAL	Grande	Privada	Industrial	Industria metalúrgica	100,00%
HALCOR	Grande	Privada	Industrial	Industria metalúrgica	96,36%
PUBLIC POWER CORPORATION (PPC)	Grande	Propiedad Pública	Agua y Energía	Suministro de energía	92,73%
NATIONAL BANK OF GREECE	Mediana	Privada	Servicios	Actividades financieras	92,73%
MOTOR OIL HELLAS	Mediana	Privada	Agua y Energía	Suministro de energía	92,73%
HELLENIC PETROLEUM	Grande	Privada	Agua y Energía	Suministro de energía	89,09%
FRIGOGLASS	Mediana	Privada	Industrial	Fabricación de bebidas	85,45%
PIRAEUS BANK	Mediana	Privada	Servicios	Actividades financieras	81,82%
GEFYRA	Pequeña	Privada	Construcción	Ingeniería civil	80,00%

Tabla 17. Clasificación de las empresas de Italia con mayor nivel de sostenibilidad. Año 2011

AÑO 2011					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR		NIVEL DE SOSTENIBILIDAD
HOLCIM ITALIA	Pequeña	Filial	Construcción	Materiales de construcción	100,00%
CNH INDUSTRIAL N.V.	Grande	Privada	Industrial	Fabricación de vehículos de motor	100,00%
FIAT CHRYSLER AUTOMOBILES	Mediana	Privada	Industrial	Fabricación de vehículos de motor	100,00%
PIRELLI	Mediana	Privada	Industrial	Fabricación de productos de caucho y plásticos	100,00%
ALTRAN ITALY	Grande	Filial	Servicios	Consultoría tecnológica e innovación	98,18%
GRUPPO HERA	Grande	Privada	Agua y Energía	Suministro de energía	94,55%
GRUPPO BNL	Grande	Privada	Servicios	Actividades financieras	92,73%
ERG (GRUPPO ERG)	Grande	Privada	Agua y Energía	Suministro de energía	90,91%
AZA SPA	Grande	Privada	Agua y Energía	Suministro de energía	87,27%
EDIPOWER	Grande	Privada	Agua y Energía	Suministro de energía	87,27%

Tabla 18. Clasificación de las empresas de Italia con mayor nivel de sostenibilidad. Año 2012

AÑO 2012					
EMPRESA	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	TIPO DE ACTIVIDAD	NIVEL DE SOSTENIBILIDAD
ACEA	Grande	Privada	Agua y Energía	Suministro de energía	100,00%
FERALPI GROUP	Mediana	Privada	Industrial	Industria metalúrgica	100,00%
HOLCIM ITALIA	Pequeña	Filial	Construcción	Materiales de construcción	100,00%
PIRELLI	Mediana	Privada	Industrial	Fabricación de productos de caucho y plásticos	100,00%
FERROVIE DELLO STATO - CORPORATE ENVIRONMENT	Grande	Propiedad del Estado	Servicios	Transporte terrestre	96,36%
GRUPPO HERA	Grande	Privada	Agua y Energía	Suministro de energía	94,55%
SALINI COSTRUTTORI	Grande	Privada	Construcción	Construcción	92,73%
TERNA	Grande	Privada	Agua y Energía	Suministro de energía	90,91%
AZA SPA	Grande	Privada	Agua y Energía	Suministro de energía	81,82%
COLCEM	Grande	Privada	Construcción	Materiales de construcción	81,82%

Tabla 19. Clasificación global de las empresas con mayor nivel de sostenibilidad. Años 2011 y 2012

AÑO 2011					AÑO 2012				
EMPRESA	PAÍS	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR	EMPRESA	PAÍS	TAMAÑO	TIPO DE ORGANIZACIÓN	SECTOR
FRAPORT AG	Alemania	Grande	Privada	Servicios	FRAPORT AG	Alemania	Grande	Privada	Servicios
MUNICH AIRPORT	Alemania	Grande	Privada	Servicios	MUNICH AIRPORT	Alemania	Grande	Privada	Servicios
SANLUCAR FRUITS	Alemania	Grande	Privada	Primario	SANLUCAR FRUITS	Alemania	Grande	Privada	Primario
REWE GROUP	Alemania	Grande	Privada	Servicios	REWE GROUP	Alemania	Grande	Privada	Servicios
TYCZKA UNTERNEHMENSGRUPPE	Alemania	Grande	Privada	Agua y Energía	MERCK GERMANY	Alemania	Mediana	Privada	Industrial
COCA-COLA DEUTSCHLAND	Alemania	Grande	Privada	Industrial	TCHIBO	Alemania	Grande	Privada	Servicios
REWE INTERNATIONAL AG	Austria	Grande	Privada	Servicios	REWE INTERNATIONAL AG	Austria	Grande	Privada	Servicios
ABEINSA	España	Grande	Privada	Agua y Energía	INA	Croacia	Grande	Privada	Agua y Energía
ABERTIS	España	Grande	Privada	Otros	ABEINSA	España	Grande	Privada	Agua y Energía
BANCO SABADELL	España	Grande	Privada	Servicios	AGBAR	España	Grande	Privada	Agua y Energía
ENDESA	España	Grande	Privada	Agua y Energía	AQUALIA	España	Grande	Privada	Agua y Energía
GRUPO ACS	España	Grande	Privada	Construcción	BANCO SABADELL	España	Grande	Privada	Servicios
GRUPO UNIÓN FENOSA GAS	España	Mediana	Privada	Agua y Energía	BANKINTER	España	Grande	Privada	Servicios
TELVENT	España	Grande	Privada	Servicios	ENDESA	España	Grande	Privada	Agua y Energía
HELLENIC CABLES	Grecia	Mediana	Privada	Industrial	RENEF	España	Grande	Propiedad Pública	Servicios
HOLCIM ITALIA	Italia	Pequeña	Filial	Construcción	HELLENIC CABLES	Grecia	Mediana	Privada	Industrial
CNH INDUSTRIAL N.V.	Italia	Grande	Privada	Industrial	ELVAL	Grecia	Grande	Privada	Industrial
FIAT CHRYSLER AUTOMOBILES	Italia	Mediana	Privada	Industrial	ACEA	Italia	Grande	Privada	Agua y Energía
PIRELLI	Italia	Mediana	Privada	Industrial	FERALPI GROUP	Italia	Mediana	Privada	Industrial
ÁGUAS DO DOURO E PAIVA	Portugal	Grande	Propiedad del Estado	Agua y Energía	HOLCIM ITALIA	Italia	Pequeña	Filial	Construcción
ANA	Portugal	Grande	Privada	Servicios	PIRELLI	Italia	Mediana	Privada	Industrial
AXA SEGUROS PORTUGAL	Portugal	Grande	Privada	Servicios	CROWN VAN GELDER	Países Bajos	Grande	Privada	Industrial
BANCO ESPIRITO SANTO	Portugal	Grande	Privada	Servicios	GRUPA LOTOS	Polonia	Grande	Propiedad del Estado	Agua y Energía
BRISA	Portugal	Grande	Privada	Servicios	ANA	Portugal	Grande	Privada	Servicios
GRUPO PORTUCEL	Portugal	Grande	Privada	Industrial	MANCHESTER AIRPORT GROUP	Reino Unido	Grande	Propiedad Pública	Servicios
RECKITT BENCKISER	Reino Unido	Mediana	Privada	Industrial	MOESK	Rusia	Grande	Filial	Agua y Energía
MOESK	Rusia	Grande	Filial	Agua y Energía	NORILSK NICKEL	Rusia	Grande	Privada	Primario
OAO NIZHNEKAMSKNEFTEKHIM	Rusia	Grande	Asociación	Industrial	OAO NIZHNEKAMSKNEFTEKHIM	Rusia	Grande	Asociación	Industrial
HOLMEN	Suecia	Grande	Privada	Industrial	HOLMEN	Suecia	Grande	Privada	Industrial
BURSAGAZ A.S.	Turquía	Pequeña	Privada	Agua y Energía	BURSAGAZ A.S.	Turquía	Pequeña	Privada	Agua y Energía

SISTEMA DE INTELIGENCIA DE NEGOCIO PARA ANALIZAR INFORMACIÓN HIDROLÓGICA.

Farid García Valentín, Yasser Vázquez Alfonso, Yerandy Carlos Rodríguez Salgado, Jorge Lázaro López & Eddy Pérez Coba.

*Empresa de Aprovechamiento Hidráulico de Mayabeque (EAHM)

** Universidad de la Habana, Facultad de Turismo.

*** Universidad Agraria de la Habana

ABSTRACT

The intelligence system of business is one of the manifestation of the contemporary technical scientist revolution widely spread in developed countries. Currently the informatic systems applied to hydrology are widely used in the international practice since they have evolved the hydrological research methodologies enabling to solve flooding matters increased associated with climate change, forecasting and modelling future sceneries. In the research an analysis of concepts, tools and methodologies have been done which are considered important in the problem, reflecting the current state of its solution. The Mayabeque Hydraulic Enterprise Use (EAHM) has among its missions and tasks the strict control of hydraulic resources in the province. It has a large volume of information that is daily generated. As a response to this necessity we decided to carry on an Intelligence System of Business for the analysis of the hydrological information the construction and implementation of this system, using the Data Base Manager PostgreSQL and the multidimensional modelling tool Pentaho, presented a good acceptance by the Enterprise board and it helps in making decisions of the board of directors at enterprise level.

KEYWORDS: Intelligence System of Business, Hydrological information, Data Mart

RESUMEN

Los Sistemas de Inteligencia de Negocio es una de las manifestaciones de la revolución científico-técnica contemporánea ampliamente generalizada en países desarrollados. Actualmente los sistemas informáticos aplicados a la hidrología son muy usados en la práctica internacional ya que han evolucionado las metodologías de investigación hidrológica, posibilitando solucionar los problemas de inundaciones cada vez más relacionado con el cambio climático, realizar pronósticos y modelar escenarios futuros. En la investigación se realiza un análisis de los conceptos, herramientas y metodologías que se consideran importantes en el problema, reflejando el estado actual en la solución del mismo. La Empresa de Aprovechamiento Hidráulico de Mayabeque (EAHM), al tener entre sus misiones y tareas el estricto control de los recursos hídricos en la provincia, posee un gran volumen de información que es generada diariamente. Como una respuesta a esa necesidad se decidió implementar un Sistema de Inteligencia de Negocio, para el análisis de la información hidrológica. La construcción e implementación de este sistema, utilizando el Gestor de Base Datos PostgreSQL y la herramienta de modelado Multidimensional Pentaho, presentó buena aceptación por parte de los directivos de la Empresa y contribuye a la toma de decisiones de los directivos a nivel empresarial.

1. INTRODUCCIÓN.

Actualmente existe la necesidad de tener una visión analítica y universal de la evolución de situaciones ambientales, sociales y administrativas a través del acceso a bases de información que se alimentan de datos de diferente naturaleza. Sin embargo, muy pocos trabajos hoy en día asumen de manera frontal los problemas de la integración de los datos del medio ambiente, en particular datos hidrológicos, considerando las necesidades de los usuarios tanto expertos como finales.

La toma de decisiones relacionada con los recursos hídricos es un área que requiere de información para conocer la situación actual del recurso en cuanto a calidad, disponibilidad y demanda, así como modelos que permitan evaluar escenarios prospectivos y simular los efectos que producirían ciertas acciones (Velázquez et. al.,2013).

En el mundo se reportan numerosas instituciones, las cuales se enfocan en la realización de aplicaciones informáticas para la toma de decisiones en el área de la hidrología entre estas se tiene: Fondo para la Protección del Agua (FONAG) con su Sistema de Información para la Gestión y Planificación de Recursos Hídricos (SIRH-CG), Centro de Investigación sobre Recursos Hídricos de la Universidad de Texas con el sistema (ArcHydro), Instituto Geológico y Minero de España con el Sistema de Información del Agua Subterránea (SIAS), Consejo del Sistema Veracruzano del Agua con su Sistema de Información Hidráulica (SIH) y el Sistema de Información Geográfica para la Hidrología (HSIG) de Paraguay (Rodríguez, 2014).

En Cuba se destacan diferentes instituciones que han realizado sistemas informáticos para la toma de decisiones de las Empresas de Aprovechamiento Hidráulico como son Instituto Superior

Politécnico José Antonio Echevarría (CUJAE), Centro de Información y Gestión Tecnológica de Santiago de Cuba y la Empresa de Aprovechamiento Hidráulico Mayabeque.

Tradicionalmente, los sistemas de información construidos para manipular este tipo de datos integran estos de manera ad-hoc en bases de datos que no le resguardan de manera segura, ni permiten soportar la explotación y el mantenimiento (actualización) transparente de los mismos.

En la Empresa de Aprovechamiento Hidráulico Mayabeque se desarrollan y mantienen sus propios sistemas de información para administrar datos sobre el comportamiento de los ríos, volúmenes de aguas y crecimiento de presas con respecto a otras condiciones (meteorológicas, aspectos sociales y de producción). Sin embargo, la misma observa, anota y analiza los datos de manera diferente y no proporciona herramientas orientadas al análisis y explotación de este tipo de información. De esta manera se hace evidente la necesidad de diseñar y construir un soporte para el apoyo al almacenamiento integrado, consulta, análisis, visualización y mantenimiento automático de datos hidrológicos, para tomar decisiones según las necesidades de la Empresa.

2. MATERIALES Y MÉTODOS

La presente investigación se realiza en el año 2013 en la empresa de aprovechamiento hidráulico Mayabeque (eahm) que se encuentra ubicada en el municipio de güines de la provincia de mayabeque.

Para el desarrollo del sistema de inteligencia de negocios (sin), se utiliza la metodología de desarrollo de kimball (2004), ya que este sistema tiene como propósito facilitar el entendimiento de la información hidrológica, proveniente de fenómenos naturales que ocurren en territorios de la provincia de mayabeque.

La metodología se basa en lo que kimball denomina ciclo de vida dimensional del negocio (business dimensional lifecycle). Este ciclo de vida del proyecto de dw, está basado en cuatro principios básicos según (morrow, 2014):

Centrarse en el negocio.

Construir una infraestructura de información adecuada.

Realizar entregas en incrementos significativos: (este principio consiste en crear el almacén de datos dw en incrementos entregables en plazos de seis a doce meses. En este punto, la metodología se parece a las metodologías ágiles de construcción de software).

Ofrecer la solución completa. En este se punto proporcionan todos los elementos necesarios para entregar valor a los usuarios de negocios, para esto ya se debe tener un almacén de datos bien diseñado, se deberán entregar herramientas de consulta ad hoc, aplicaciones para informes, análisis avanzado, capacitación, soporte, sitio web y documentación.

En la figura 1 se presenta el ciclo de vida de la metodología de Kimball que consiste en:

Planificación del Proyecto: En este proceso se determina el propósito del proyecto de DW/BI, sus objetivos específicos y el alcance del mismo, los principales riesgos y una aproximación inicial a las necesidades de información.

Definición de Requerimientos del Negocio: La definición de requerimientos, es un proceso en el que se entrevista al personal de negocio y técnico, aunque siempre conviene, tener un poco de preparación previa. En esta tarea, se debe aprender sobre el negocio, los competidores, la industria y los clientes del mismo. Se debe dar una revisión a todos los informes posibles de la organización; rastrear los documentos de estrategia interna; entrevistar a los empleados, analizar lo que se dice en la prensa acerca de la organización, la competencia y la industria y se deben conocer los términos y la terminología del negocio.

Modelado Dimensional: Es un proceso dinámico y altamente iterativo. Comienza con un modelo dimensional de alto nivel obtenido a partir de los procesos priorizados y descritos en la tarea anterior.

Diseño Físico: En esta tarea, se contestan las siguientes preguntas:

¿Cómo puede determinar cuán grande será el sistema de DW/BI?

¿Cuáles son los factores de uso que llevarán a una configuración más grande y más compleja?

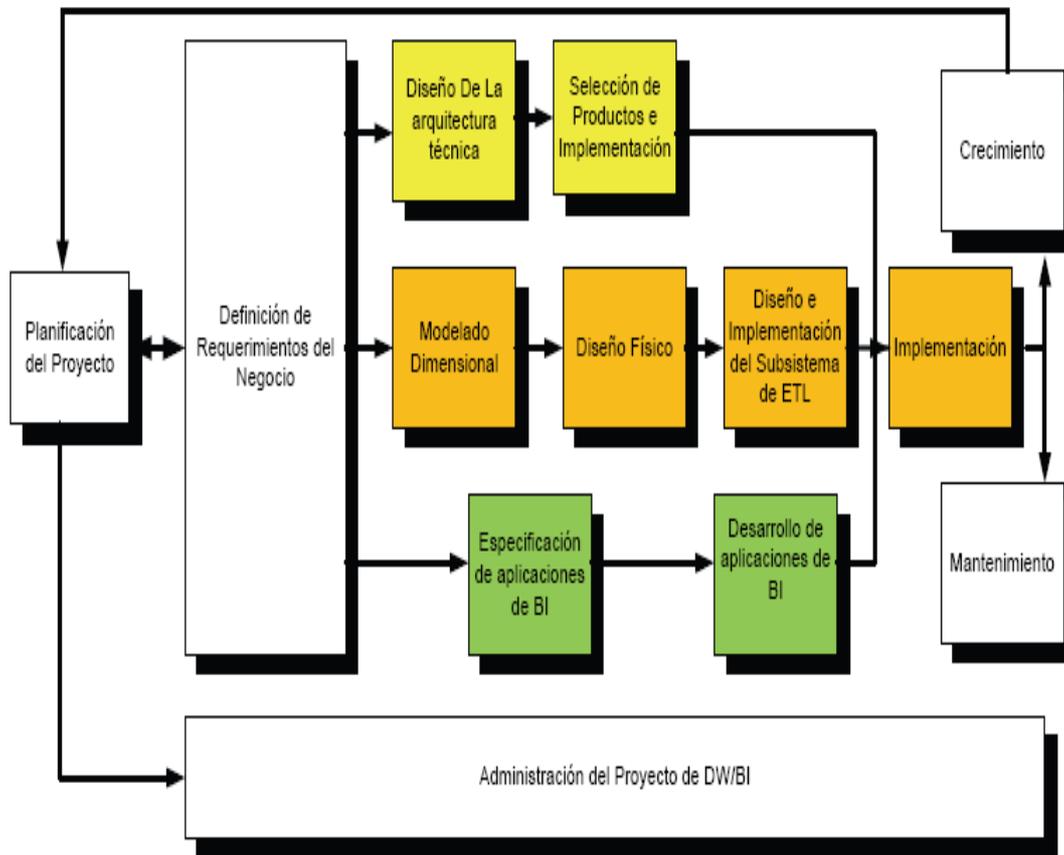


Figura 1: Ciclo de vida, metodología Kimball.
Fuente: Morrow(2014)

¿Cómo se debe configurar el sistema?

¿Cuánta memoria y servidores se necesitan? ¿Qué tipo de almacenamiento y procesadores?

¿Cómo instalar el software en los servidores de desarrollo, prueba y producción?

¿Qué necesitan instalar los diferentes miembros del equipo de DW/BI en sus estaciones de trabajo?

¿Cómo convertir el modelo de datos lógico en un modelo de datos físicos en la base de datos relacional?

¿Cómo conseguir un plan de indexación inicial?

¿Debe usarse la partición en las tablas relacionales?

Diseño e Implementación del subsistema de Extracción, Transformación y Carga (ETC):

El subsistema ETC es la base sobre la cual se alimenta el DW. Si se diseña adecuadamente, puede extraer los datos de los sistemas de origen de datos, aplicar diferentes reglas para aumentar la calidad y consistencia de los mismos, consolidar la información proveniente de distintos sistemas, y finalmente cargar (grabar) la información en el DW en un formato acorde para la utilización por parte de las herramientas de análisis.

Implementación: La implementación representa la convergencia de la tecnología, los datos y las aplicaciones de usuarios finales accesible desde el escritorio del usuario del negocio. Existen varios factores extras que aseguran el correcto funcionamiento de todas estas piezas, entre ellos se encuentran la capacitación, el soporte técnico y la comunicación.

Mantenimiento y Crecimiento del DW: Para administrar el entorno del DW existente es importante enfocarse en los usuarios de negocio, los cuales son el motivo de su existencia, además de gestionar adecuadamente las operaciones del DW, medir y proyectar su éxito y comunicarse constantemente con los usuarios para establecer un flujo de retroalimentación. En esto consiste el mantenimiento. Finalmente, es importante sentar las bases para el crecimiento y evolución del DW en donde el aspecto clave es manejar el crecimiento y evolución de forma iterativa utilizando el ciclo de vida propuesto, y establecer las oportunidades de crecimiento y evolución en orden por nivel de prioridad.

Para definir los requerimientos del negocio se hace necesario realizar entrevistas al personal, para concretar las funciones del SIN que se actualizará en el departamento de desarrollo de la EAHM.

Requerimientos generales: En el departamento de desarrollo de la EAHM se registra un gran conjunto de información generada diariamente, proveniente de fenómenos meteorológicos que ocurren en la provincia. Al implementarse el SIN se sustituirá una versión anterior de un DM, que posibilitará centralizar y organizar la información, logrando así, que su búsqueda sea más fácil y que se pueda tener una mejor visualización de los datos.

El software constará con herramientas como gráficos y reportes que no se encontraban en la versión anterior y son de gran importancia para facilitar tareas al directivo responsable de la toma de decisiones.

Requerimientos arquitectónicos: Los requerimientos arquitectónicos son herramientas informáticas necesarias para el desarrollo del SIN que requiere:

Un SGBD para almacenar la información que se generará del proceso de ETC. Se organizará la información en esquemas, el esquema pntemp contendrá las tablas históricas y por otra parte en el esquema dm_idro se almacenarán las tablas de hechos y de dimensiones.

El sistema podrá ser utilizado bajo los sistemas operativos Windows o Linux.

Se utilizará Kettle Pentaho para realizar el proceso de ETC, en dicho proceso se capturará toda la información de diferentes fuentes logrando una centralización de la información.

Se utiliza Pentaho BI Server, pues en él se implementará el sistema y sus funcionalidades.

Se debe utilizar Java JDK 1.5 o una versión superior ya que la mayoría de las herramientas a utilizar están implementadas en Java.

Requerimientos de desarrollo: Para el desarrollo del SIN los requerimientos se dividen en dos grupos:

Hardware, donde se necesita una computadora con los siguientes requisitos mínimos:

Procesador 1.50 GHz.

1GB de RAM.

Espacio en disco duro 20 Gb.

Lector de DVD.

Software. Sistema Operativo Windows o Linux.

Navegador web.

Máquina virtual de Java.

Plataforma de Pentaho.

3. RESULTADO Y DISCUSIÓN

El sistema se implementó en ambiente web sobre el servidor de Inteligencia de Negocio Pentaho BI Server. La primera interfaz que muestra el sistema es donde el usuario tendrá que escribir su usuario y contraseña, insertados previamente por el administrador del sistema. Esta interfaz es generada por el servidor, y a diferencia del software anterior cada usuario posee contraseña propia.

A continuación se muestra la interfaz de bienvenida al sistema de la EAHM, en el mismo se inserta nombre del usuario y contraseña dando acceso a la página principal creada por la empresa.



Figura 2: Página de autenticarse del sistema

Al encontrarse en la página principal que se muestra a continuación, se selecciona la pestaña de **S. Técnico** (Análisis Técnico) que dará inicio al SIN y ahí se muestran todas las funcionalidades que se implantaron (Figura 3).



Figura 3: Página principal de la EAHM.

En la figura 4 se muestra la salida del análisis técnico realizado a las diferentes variables que componen el sistema hidrológico de la Empresa de Aprovechamiento Hidráulico de Mayabeque.



Figura 4: Análisis Técnico

Estos resultados complementan las investigaciones realizadas en el campo del análisis hidrológico como son Nuñez (2009), FICH (2010), García y Vega (2012), las cuales fueron un punto de partida para esta investigación.

4. CONCLUSIONES

En la investigación se arriba a las siguientes conclusiones:

Se identificaron los procesos y flujos de información que interviene en el análisis de los datos hidrológicos, para dar paso a la confección del proceso de ETC, el cual permite conocer como fluye el almacenamiento de la información en las fuentes de datos para obtener las ETC, en esta se modifica la información que se muestra a los usuarios.

Se seleccionaron los métodos y herramientas de desarrollo enfocados hacia la elaboración del SIN. Estas permitieron confeccionar el sistema resultante con calidad, de acuerdo a los requerimientos de usuarios.

Se diseñaron las bases de datos relacionadas con la información hidrológica proveniente de los procesos de ETC utilizando los gestores de base de datos. Con esta se lograron almacenar una cierta cantidad de datos históricos procedentes de las lluvias acumuladas a través de los años.

Se implementó un Sistema de Inteligencia de Negocio para analizar la información hidrológica en la EAHM, teniendo como herramientas incorporadas los gráficos o dashboards y reportes

REFERENCIAS

- [1] FICH (2010): Sistemas de alerta y pronóstico hidrológico. Disponible en: <http://fich.unl.edu.ar/pagina/sistemas-alerta-pronostico-hidrologico/358/>. [Citado el: 28 de Mayo de 2014.].
- [2] García ,AF y Vega, J. L (2012): Desarrollo de un Data Mart para la Empresa de Aprovechamiento Hidráulico Mayabeque. Mayabeque: s.n Tesis presentada en opción al Título de Ingeniero en Informática. Universidad Agraria de la Habana. Provincia Mayabeque.
- [3] Kimball, R. (2004): The Data Warehouse ETL Toolkit. s.l. : Wiley Publishing, Inc. Chichester.
- [4] Núñez, G. (2009): Lagos, lagunas y embalses. Volúmenes característicos de un embalse. Disponible en: <http://lagoslagunasyembalses.blogspot.com/2009/11/volumenes-caracteristicos-de-un-embalse.html>. 05 de Noviembre de 2009. [Citado el: 11 de Abril de 2014.].

- [5] Morrow, J (2014): . Inteligencia de Negocio. Metodología de Kimball. Disponible en: <http://inteligenciadenegociosval.blogspot.com/2014/01/metodologia-de-kimball.html> [Citado el: 13 de 04 de 2014,].
- [6] Rodríguez,Y (2014): . Sistema de Inteligencia de Negocio para analizar la información hidrológica. Tesis presentada en opción al Título de Ingeniero en Informática.Universidad Agraria de la Habana. Provincia Mayabeque.
- [7] Velázquez, J et. al. (2013): Implementación del modelo de datos ArcHydro en la región hidrológica Balsas. Disponible en http://www.inecc.gob.mx/descargas/cuencas/cong.../13_jaime_velazquez.pdf.

CRECIMIENTO ECONÓMICO, MIGRACIÓN Y REMESAS EN MÉXICO

Ignacio Caamal Cauich*, Verna Gricel Pat Fernández*, David Martínez Luis**, José Félix García Rodríguez** & Alberto Pérez Fernández***

* Universidad de Chapingo, México

** Universidad Juárez Autónoma de Tabasco. México

*** Universidad Autónoma del Carmen, Campeche, México

ABSTRACT

Economic growth is directly related to the behavior of the volume and value of the final goods and services produced in a given period, gross domestic product. Economic stagnation manifests itself in low rates of growth of the gross domestic product and results in the reduction of opportunities of employment and low wages and is one of the factors that explain the Mexican population migration to other countries, especially to the United States of America. Low rates of growth of the Mexican economy, unemployment and low wages are the factors that explain the migration of the population in Mexico.

KEYWORDS: poverty, GDP, GDP per capita.

RESUMEN

El crecimiento económico está directamente relacionado con el comportamiento del volumen y valor real de los bienes y servicios finales producidos, en un periodo determinado, producto interno bruto. El estancamiento económico se manifiesta en las bajas tasas de crecimiento del producto interno bruto y provoca la reducción de las oportunidades de empleo y bajos salarios y es uno de los factores que explican la migración de la población mexicana a otros países, especialmente a los Estados Unidos de América. Las bajas tasas de crecimiento de la economía mexicana, el desempleo y los bajos salarios son los factores que explican la migración de la población en México.

1. INTRODUCCIÓN

La migración está directamente relacionada con el nivel de crecimiento económico de un país. El crecimiento económico se mide con el comportamiento del producto interno bruto y el desarrollo económico a través de los niveles de bienestar. El producto interno bruto se refiere al valor total de los bienes y servicios finales producidos en un periodo.

El crecimiento económico de México en las últimas décadas ha sido bajo, lo que refleja la existencia de un estancamiento económico, que genera desempleo, subempleo, bajos salarios y migración. El estancamiento económico que se observa en México y en la mayoría de los países en vías de desarrollo de Latinoamérica ha impulsado la migración de los países y regiones de menor desarrollo a los países y regiones de mayor desarrollo.

En este contexto, el objetivo de este trabajo es caracterizar el comportamiento del crecimiento económico, la migración y las remesas en México. Los apartados del trabajo son crecimiento económico, migración y remesas. La base de este trabajo se encuentra en los capítulos de libro publicados, los cuales son *Pobreza, crecimiento económico, desempleo y salarios en México* (Caamal et al., 2014) y *Pobreza, migración y remesas en México* (Caamal y Pat, 2015). El aporte de este trabajo se encuentra en la actualización de los datos y la ampliación de la discusión sobre los problemas del crecimiento económico, la migración y las remesas.

2. CRECIMIENTO ECONÓMICO

2.1. Concepto

El crecimiento económico se refiere al incremento en el tiempo de la cantidad, gama, y valor de los bienes y servicios producidos en la economía. Existe crecimiento económico cuando la cantidad y valor real de los bienes y servicios producidos de un periodo a otro se incrementa. El crecimiento económico se mide a través del producto interno bruto real, que se refiere al valor total real de los bienes y servicios producidos de un periodo a otro periodo (Sánchez, 2009).

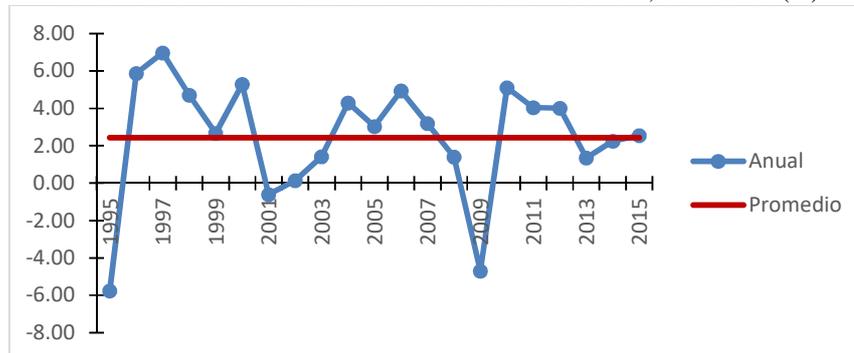
El crecimiento económico está directamente relacionado con las inversiones en capital físico, en capital humano, en conocimiento, en innovación tecnológica, entre otras inversiones. Así como con las políticas públicas de inversión en infraestructura, en financiamiento, en educación e investigación, en estímulos, entre otras. El crecimiento económico es una condición necesaria para el desarrollo económico, el cual consiste en un conjunto de cambios estructurales, mejor distribución de la riqueza generada y un mayor bienestar, mayor cantidad y calidad de bienes y servicios obtenidos, para la sociedad (Zermeño, 2004).

2.2. Producto interno bruto y población

2.2.1. Comportamiento del producto interno bruto

Al considerar la tasa de crecimiento del PIB en términos reales como indicador del crecimiento económico, se tiene que durante el periodo de 1994 al 2015, el crecimiento económico de México se incrementó en 65.8%, con una tasa de crecimiento promedio anual de 2.4% (gráfica 1).

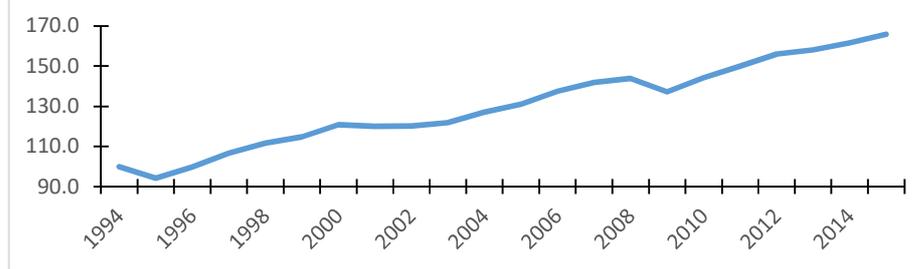
Gráfica 1. Tasa de crecimiento anual del PIB de México, 1995-2015 (%).



Fuente: Elaborado con datos del Banco Mundial, 2016.

A pesar de que, durante todo el periodo de estudio, el crecimiento económico promedio de México ha sido positivo, en algunos años ha presentado disminuciones, inclusive menores a cero, tal es el caso de 1995 donde disminuyó 5.76% respecto al año anterior, para el año 2001 disminuyó en 0.6% y en 2009 en 4.7% (gráfica 1). Esas disminuciones coinciden con factores internos como la crisis de 1994 y con factores externos como la desaceleración de la economía de Estados Unidos en el 2001 y la crisis económica mundial que comenzó en el año 2007.

Gráfica 2. Tasa de crecimiento acumulada anual del PIB de México, 1994-2015 (%).



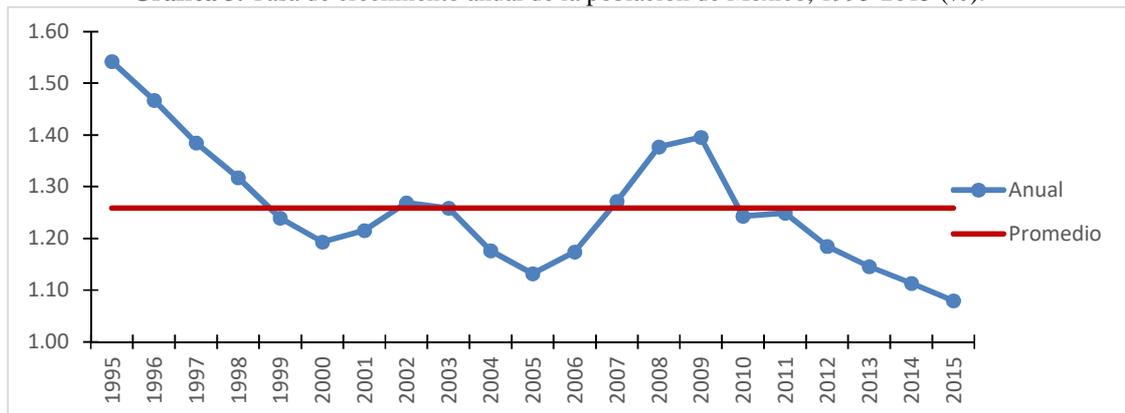
Fuente: Elaborado con datos del Banco Mundial, 2016.

En términos generales, el incremento de la tasa de crecimiento del PIB acumulado presenta una tendencia creciente, tal como se puede observar en la gráfica 2. Lo cual indica que la economía de México ha crecido durante el periodo de 1994 al 2015, con decrecimientos y una tasa de crecimiento promedio anual baja.

2.2.2. Comportamiento de la población

Aunque, durante el periodo de estudio de 1994 a 2015, la población de México se ha incrementado en 30.0%, a una tasa de crecimiento promedio anual de 1.3%, la tasa de crecimiento de la población presenta una tendencia al decrecimiento, al pasar de una tasa de crecimiento de 1.54% de 1994 a 1995, a 1.08% del 2014 al 2015, tal como se observa en la gráfica 3.

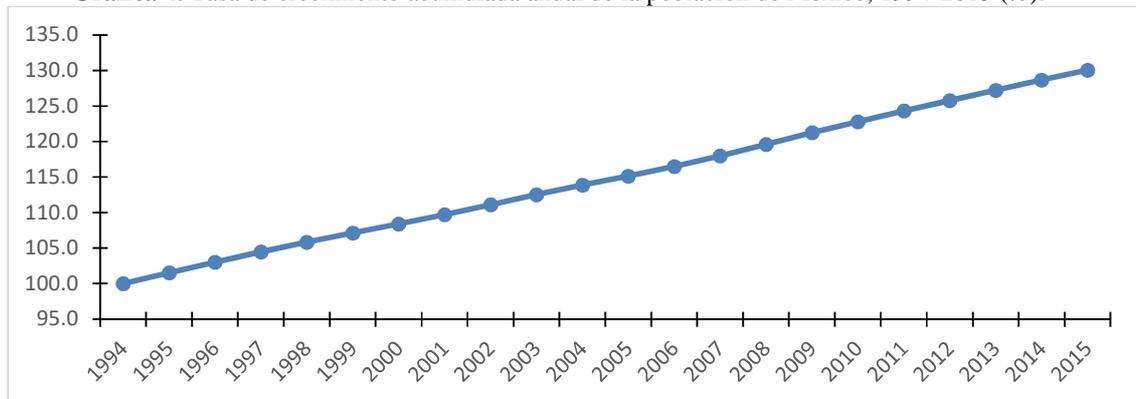
Gráfica 3. Tasa de crecimiento anual de la población de México, 1995-2015 (%).



Fuente: Elaborado con datos del CONAPO (2016b).

Los resultados reflejan que la tasa de crecimiento promedio anual del PIB fue mayor que la tasa de crecimiento promedio anual de la población durante el periodo estudiado, de 1994 a 2015.

Gráfica 4. Tasa de crecimiento acumulada anual de la población de México, 1994-2015 (%).



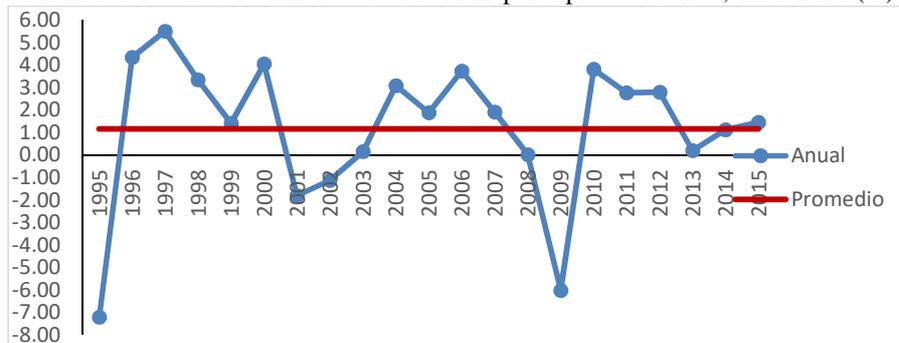
Fuente: Elaborado con datos del CONAPO (2016b).

La tasa de crecimiento de la población en términos acumulados, muestra que la población se ha incrementado durante el periodo de 1994 al 2015 en forma constante, menores a la tasa de crecimiento del producto interno bruto durante el mismo periodo (gráfica 4).

2.2.3. Comportamiento del PIB per cápita

El PIB per cápita de México durante el periodo de 1994-2015 se incrementó en 27.5%, a una tasa promedio anual de alrededor de 1.2%, presentando diferentes variaciones durante el periodo de estudio, tal como se puede observar en la gráfica 5, explicada por el comportamiento del crecimiento anual del PIB y de la población.

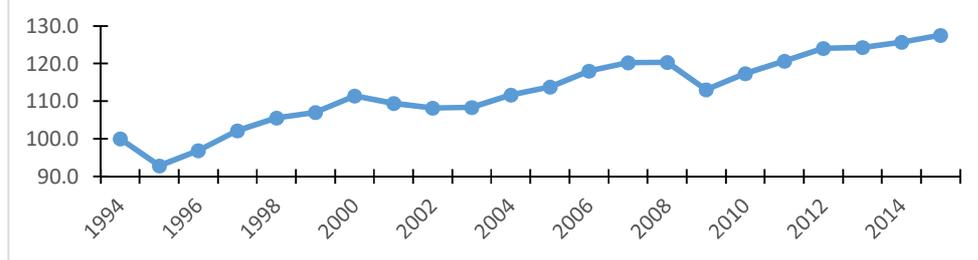
Gráfica 5. Tasa de crecimiento anual del PIB per cápita de México, 1995-2015 (%).



Fuente: Elaborado con datos del Banco Mundial (2016) y CONAPO (2016b).

El comportamiento de la tasa de crecimiento acumulada del producto interno bruto per cápita está directamente relacionada con las tasas de crecimiento global del producto interno bruto y de la población, lo que explica las oscilaciones que se observan en éste.

Gráfica 6. Tasa de crecimiento acumulada anual del PIB per cápita de México, 1994-2015 (%).

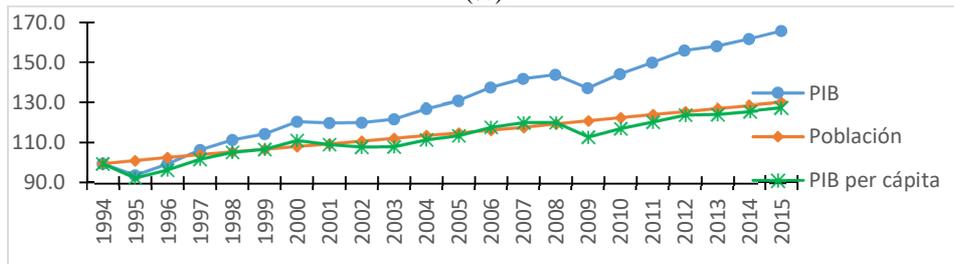


Fuente: Elaborado con datos del Banco Mundial (2016) y CONAPO (2016b).

En términos acumulados, se puede observar que el PIB per cápita de México se ha incrementado lentamente en el periodo estudiado (gráfica 6), muy por debajo del crecimiento que han tenido otros países en vías de desarrollo, el cual se explica por las bajas tasas de crecimiento del producto interno bruto per cápita y las tasas de crecimiento de la población.

2.2.4. Comportamiento del PIB, población y PIB per cápita

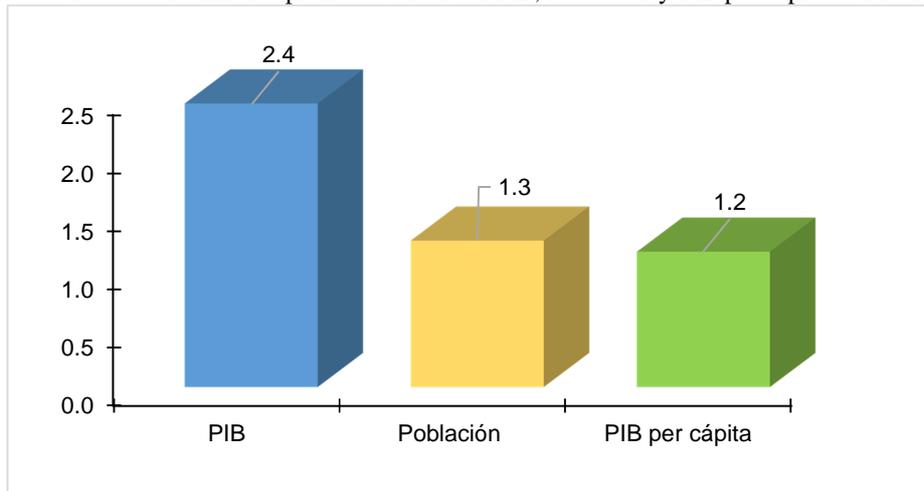
Gráfica 7. Tasas de crecimiento acumuladas anual del PIB, población y PIB per cápita de México, 1994-2015 (%).



Fuente: Elaborado con datos del Banco Mundial (2016) y CONAPO (2016b).

Las tasas de crecimiento acumuladas del PIB y de la población reflejan un lento crecimiento de la brecha entre ambos indicadores, que explica básicamente las bajas tasas de crecimiento del PIB per cápita (gráfica 7).

Gráfica 8. Tasa de crecimiento promedio anual del PIB, Población y PIB per cápita de México (%).



Fuente: Elaborado con datos del Banco Mundial (2016) y CONAPO (2016b).

Al comparar las tasas de crecimiento promedio anual del PIB, población y PIB per cápita durante el periodo analizado, se observa que el PIB presentó el mayor incremento (2.4%), casi del doble que el crecimiento de la población (1.3%) y el PIB per cápita (1.2%) como se puede apreciar en la gráfica 8.

3. MIGRACIÓN

3.1. Concepto

La migración es el desplazamiento de personas que cambian su residencia habitual desde una unidad político-administrativa hacia otra dentro de un mismo país, o que se mueven de un país a otro, en un periodo determinado (CONAPO, 2016a).

El decrecimiento y el estancamiento económico es uno de los factores determinantes por las que la población tiende a migrar de residencia para conseguir mejores condiciones de vida para su familia, que en su país de origen se les dificulta obtener, la migración se mueve de las regiones menos desarrolladas a las más desarrolladas, es por eso que la mayoría de los mexicanos que migran, lo hacen a Estados Unidos de América (EUA), por su cercanía y nivel de desarrollo económico, para obtener una mejor calidad de vida.

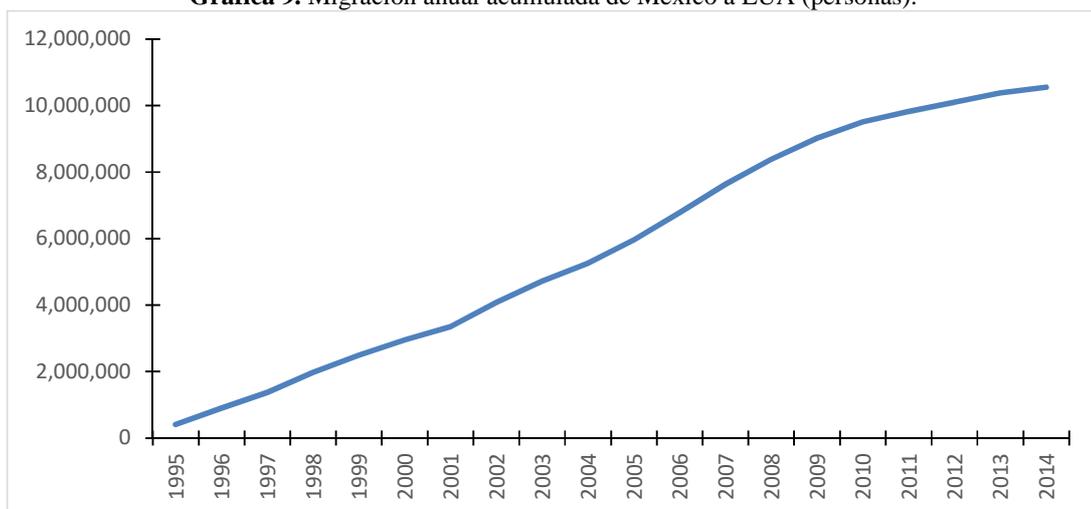
La migración de México a EUA ha pasado por varias etapas y toma relevancia en el siglo XX por dos razones fundamentales, el incremento en la demanda de trabajo de EUA, debido a las necesidades generadas durante las guerras mundiales, y al estancamiento económico de México en las últimas décadas, que refleja las limitaciones del modelo económico neoliberal, con bajos niveles de crecimiento y altos niveles de desempleo y subempleo.

La migración actual en México tiene una gran importancia, debido a que: 1. Refleja las limitaciones de la política económica de México para impulsar el crecimiento económico, crear empleos suficientes, salarios bien remunerados y generar bienestar para la mayoría de la población; 2. Genera una cantidad importante de divisas que tiene un impacto positivo en la balanza de pagos y en la reducción de la pobreza, especialmente en el sector rural; y 3. Se ha convertido en una de las fuentes más importantes de ingreso de divisas, que permite resolver problemas de déficit de la balanza comercial y de pagos.

3.2. Comportamiento de la migración

La migración en los países receptores tiene un impacto económico positivo ya que los migrantes se encuentran en edad productiva y son generadores de valor, al mismo tiempo en los países de origen, la migración también tiene un efecto positivo al reducir el subempleo y desempleo y aumentar la remisión de divisas al país, la cual favorece a la balanza de pagos y la reducción de la pobreza.

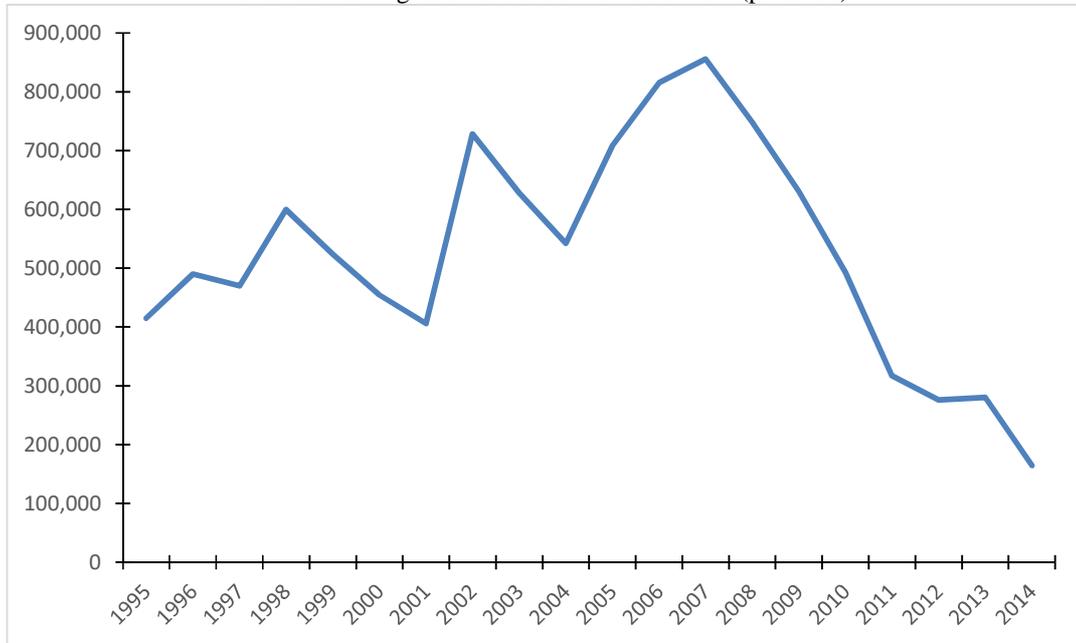
Gráfica 9. Migración anual acumulada de México a EUA (personas).



Fuente: Elaborado con datos de STPS (2016) y Passel *et al.* (2012).

En México la migración ha venido disminuyendo en los últimos años, ya que de los 415,054 migrantes reportados en 1995, estos descendieron a 164,621 en 2014, lo que representó una disminución del 60.3% en ese año (gráfica 10).

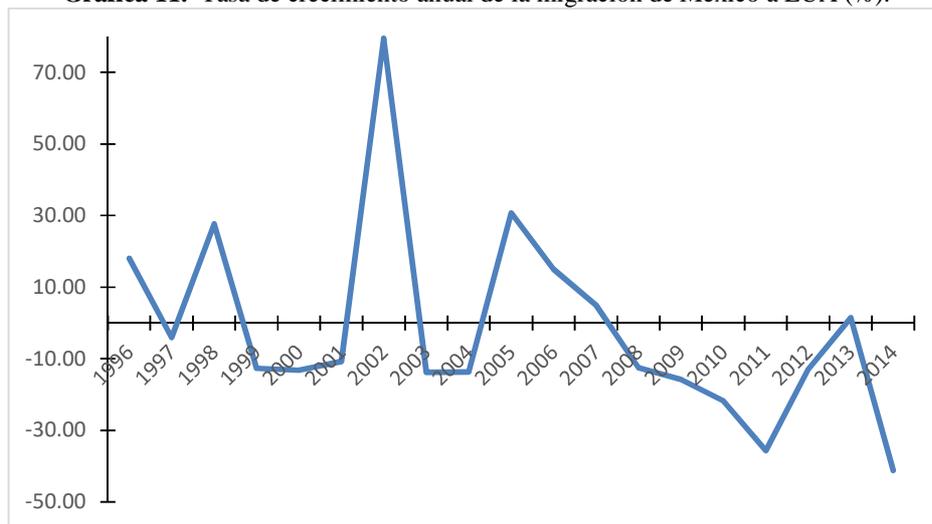
Gráfica 10. Migración anual de México a EUA (personas).



Fuente: Elaborado con datos de STPS (2016) y Passel *et al.* (2012).

Esta tendencia decreciente de la migración de trabajadores se explica por las fuertes restricciones impuestas por Estados Unidos de América y no por la recuperación económica de México.

Gráfica 11. Tasa de crecimiento anual de la migración de México a EUA (%).

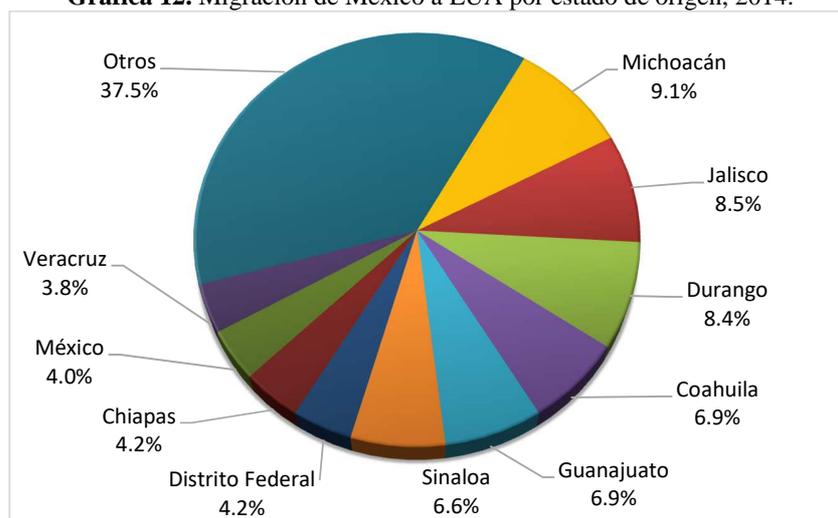


Fuente: Elaborado con datos de STPS (2016) y Passel *et al.* (2012).

3.3. Distribución de la migración

El fenómeno de la migración hacia Estados Unidos de América afecta a las 32 entidades federativas de México, siendo los estados de Michoacán, Jalisco, Durango, Coahuila, Guanajuato y Sinaloa los que presentan los más altos índices de migración, en conjunto representaron el 46.3% de migrantes reportados durante el año 2014 (gráfica 12).

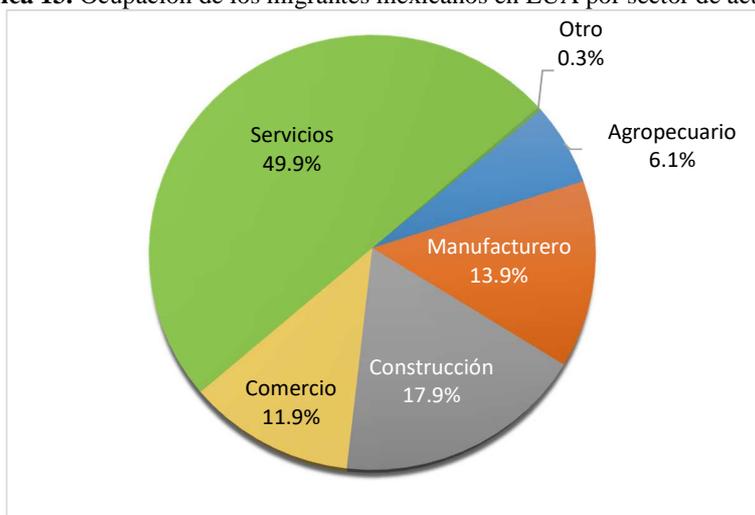
Gráfica 12. Migración de México a EUA por estado de origen, 2014.



Fuente: Elaborado con datos de STPS (2016).

Tradicionalmente los migrantes mexicanos en EUA se ocupaban en actividades ligadas a la agricultura. En la actualidad, se observa que gran parte de los migrantes se ocupan en labores del sector de los servicios y en la construcción, 67.98% de la población mexicana ocupada en EUA. Así mismo, el sector manufacturero es un importante empleador de los mexicanos residentes en Estados Unidos, ya que el 13.9% del total de migrantes se ocupa en esa área. Estos datos reflejan que, pese al bajo grado educativo de los migrantes (59% tiene menos de nueve años de escolaridad), la mayor proporción de trabajadores se emplean en sectores modernos de la economía (gráfica 13).

Gráfica 13. Ocupación de los migrantes mexicanos en EUA por sector de actividad.



Fuente: Elaborado con datos de Statista (2016).

La migración ha reducido las presiones sociales para el gobierno de México en los últimos años, puesto que los ingresos que generan las remesas (desde Estados Unidos de América, Canadá, entre otros) han reducido la situación de pobreza y desigualdad que las acciones de política no han logrado resolver. Si se toma en cuenta el ingreso medio de los hogares en el campo, en algunas regiones como el noreste del país las remesas representan hasta el 20% de los ingresos familiares.

4. REMESAS

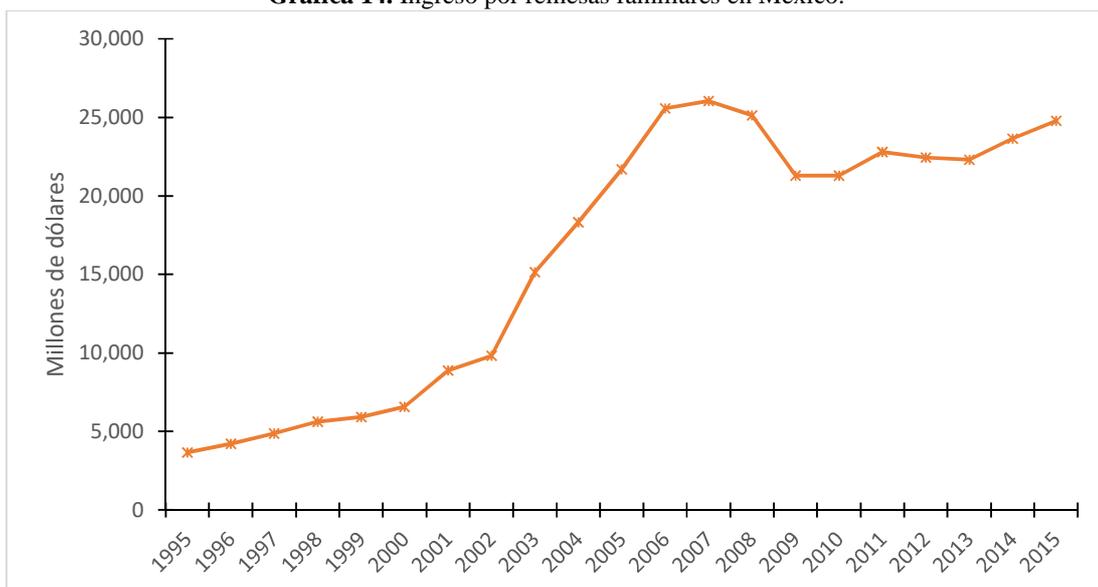
3.4. Concepto

Las remesas se refieren a las divisas que los migrantes mexicanos en otros países, especialmente, en Estados Unidos de América, envían a México. Las remesas se han convertido en las últimas décadas en un flujo de divisas de suma importancia para la economía mexicana. Este flujo constituye uno de los principales rubros en el renglón de las transferencias corrientes de la balanza de pagos y funge como una fuente de inyección de recursos en sectores específicos de la economía nacional.

3.5. Comportamiento de las remesas

El comportamiento de las remesas ha sido creciente de 1995 a 2016, alcanzando los mayores niveles de 2006 a 2008.

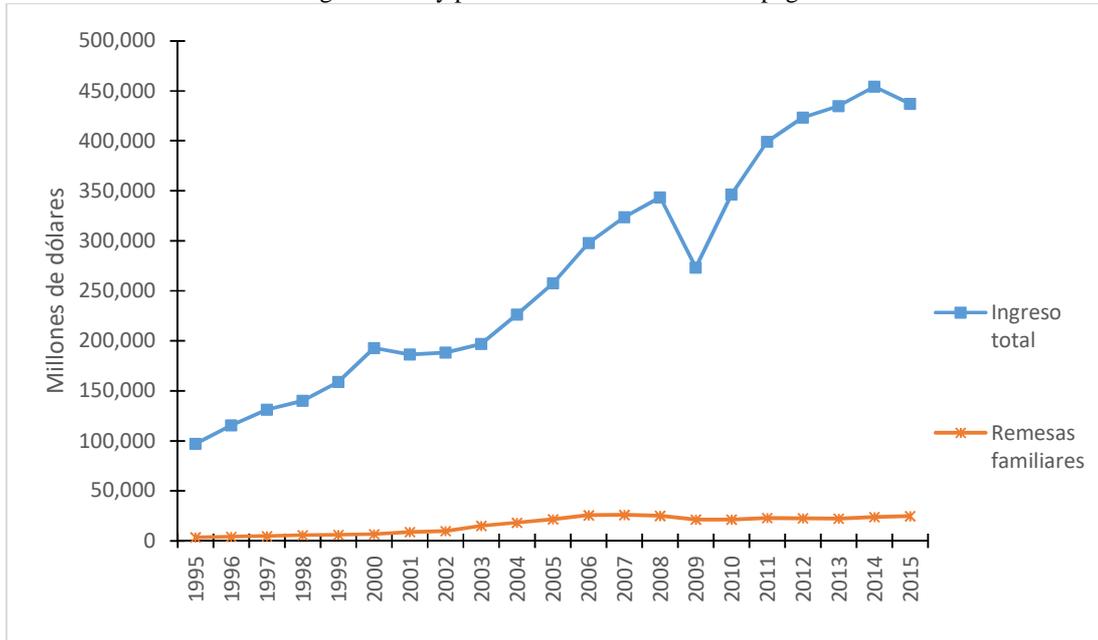
Gráfica 14. Ingreso por remesas familiares en México.



Fuente: Elaborado con datos de INEGI (2016b).

De hecho, el monto recibido en 2006 por este concepto, que fue de 25,566.8 millones de dólares, contribuyó a reducir el déficit de la cuenta corriente de la balanza de pagos en alrededor de 8.6 por ciento, lo que revela la creciente importancia de la migración como fuente de divisas.

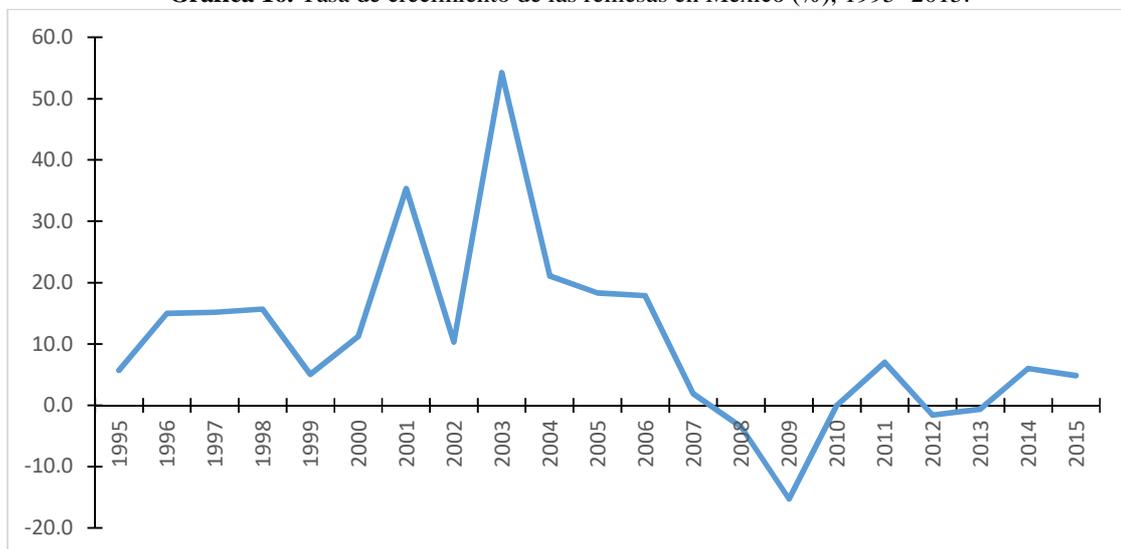
Gráfica 15. Ingreso total y por remesas de la balanza de pagos de México.



Fuente: Elaborado con datos de INEGI (2016b).

Las remesas alcanzaron sus mayores niveles en el periodo de 2006 y 2008, posteriormente con la desaceleración y crisis de la economía de Estados Unidos de América éstas empezaron a disminuir, lo que refleja una alta dependencia de la economía mexicana hacia la economía norteamericana.

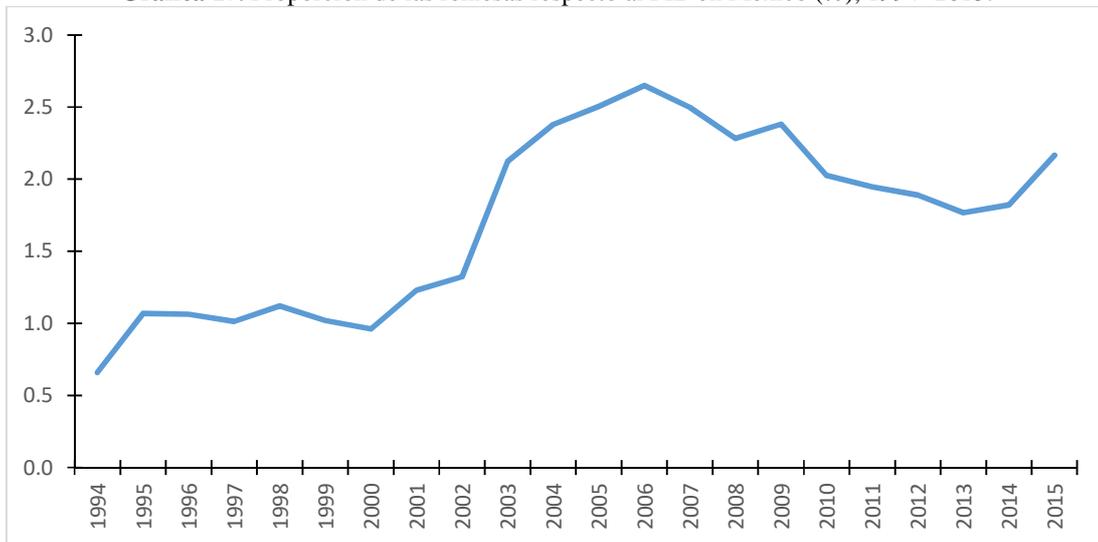
Gráfica 16. Tasa de crecimiento de las remesas en México (%), 1995- 2015.



Fuente: Elaborado con datos de INEGI (2016b).

De acuerdo con datos del Banco de México (2016), las cifras de remesas para el año 2006 ascendieron a 25,567 millones de dólares, año en que se recibió la mayor cantidad de remesas, lo que implicó un incremento de 3,879 millones y de 17.9 por ciento con respecto al 2005.

Gráfica 17. Proporción de las remesas respecto al PIB en México (%), 1994- 2015.



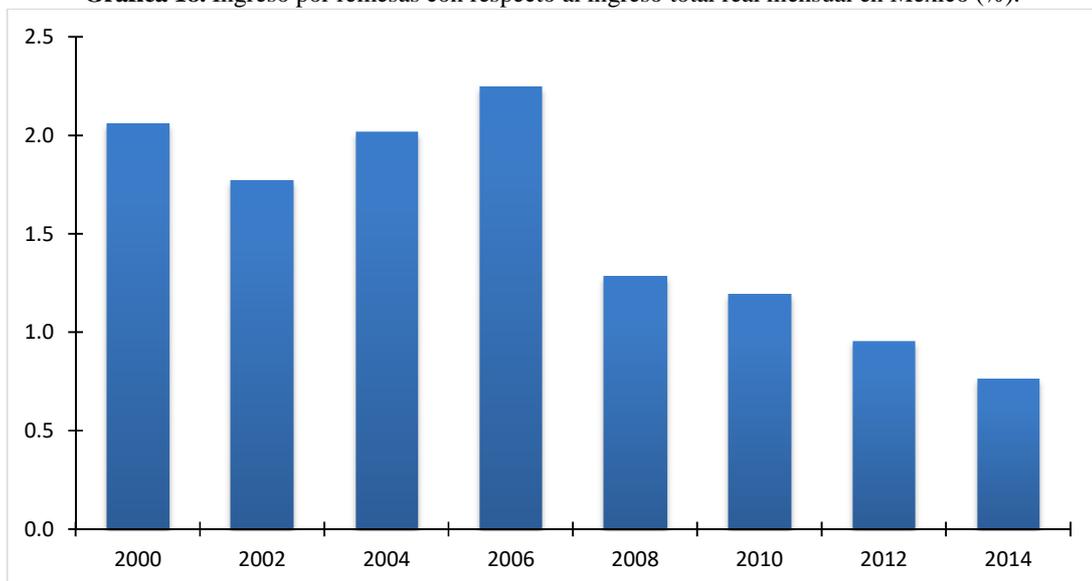
Fuente: Elaborado con datos de INEGI (2016b) y Banco Mundial (2016).

El monto de remesas en 2006 fue equivalente a 2.6 puntos porcentuales del producto interno bruto, mientras que en el 2015, dicha proporción correspondió al 2.2 por ciento. El aumento de las remesas se debe en gran medida a los aumentos de la migración laboral y la disminución, al decrecimiento de la migración debido a las restricciones impuestas por Estados Unidos de América en los últimos años.

3.6. Impacto de las remesas

Las remesas tienen un impacto en el ingreso total de los hogares en México, el cual correspondió aproximadamente al 2% en el periodo del 2000 al 2010 (gráfica 18).

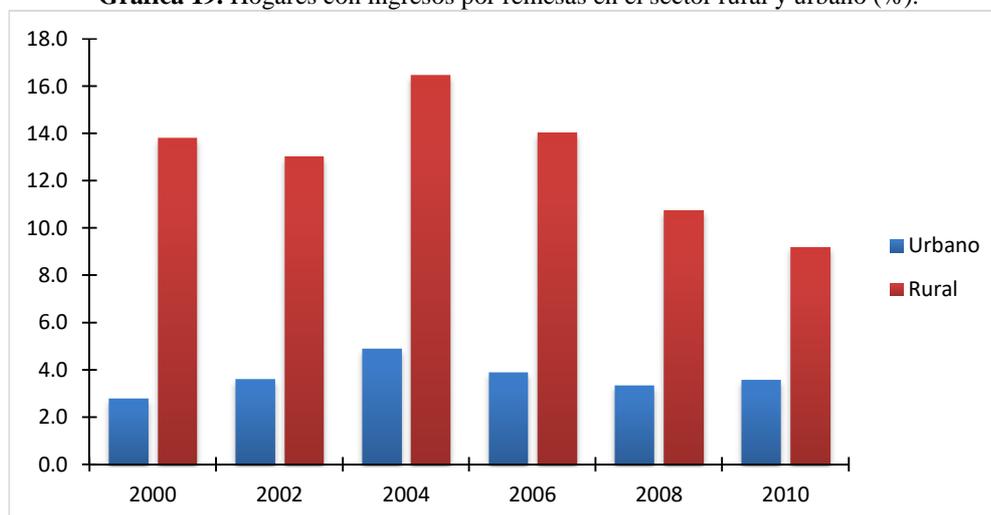
Gráfica 18. Ingreso por remesas con respecto al ingreso total real mensual en México (%).



Fuente: Elaborado con datos de INEGI (2016a).

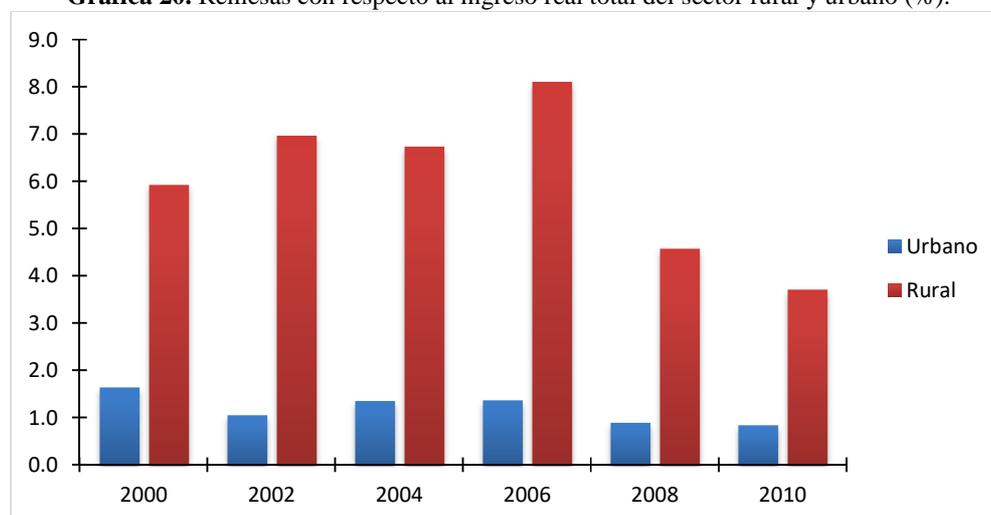
Por otro lado, dependiendo del sector en donde se ubica la población (población urbana y población rural) el impacto de las remesas varía, mientras que en el sector urbano su peso es de alrededor del 1% del ingreso total, en el sector rural es de aproximadamente 6% (gráfica 20).

Gráfica 19. Hogares con ingresos por remesas en el sector rural y urbano (%).



Fuente: Elaborado con datos de INEGI (2016a).

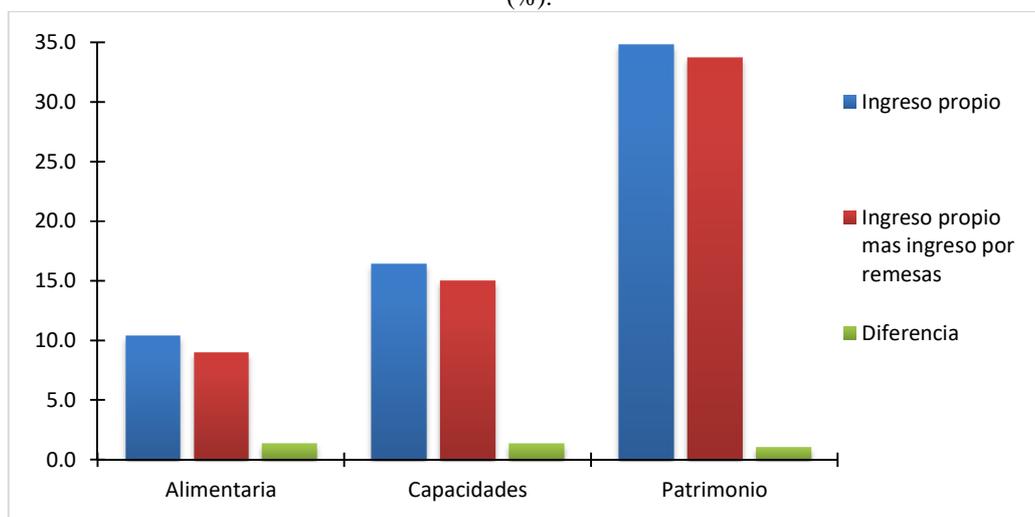
Gráfica 20. Remesas con respecto al ingreso real total del sector rural y urbano (%).



Fuente: Elaborado con datos de INEGI (2016a).

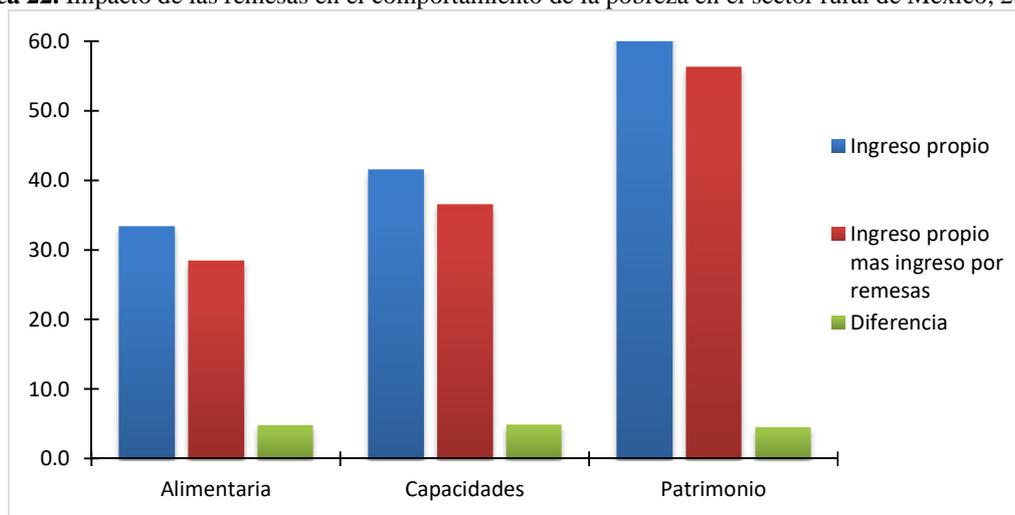
Por último, las remesas reducen la cantidad y proporción de la población en condiciones de pobreza, debido que a los ingresos obtenidos en México se le agregan las remesas que vienen del exterior, las cuales aumentan los ingresos totales y con ello rebasan los límites de la línea de pobreza. El impacto varía dependiendo del sector en donde se ubica la población, mientras que en el sector urbano el impacto es ligeramente superior al 1%, en el sector rural es de aproximadamente el 5% (gráficas 21 y 22).

Gráfica 21. Impacto de las remesas en el comportamiento de la pobreza en el sector urbano de México, 2006 (%).



Fuente: Elaborado con datos de INEGI (2016a).

Gráfica 22. Impacto de las remesas en el comportamiento de la pobreza en el sector rural de México, 2006 (%).



Fuente: Elaborado con datos de INEGI (2016a).

4. CONCLUSIONES

Los bajos niveles de crecimiento de la economía, los altos niveles de desempleo y subempleo, los bajos niveles salariales y los bajos niveles de ingreso de la población son los factores que explican y determinan la migración de la población hacia Estados Unidos de América, especialmente, en busca de oportunidades de empleo e ingresos, las cuales han estado restringidas por las bajas tasas de crecimiento de la economía y los bajos niveles salariales.

Las bajas tasas de crecimiento de la economía mexicana han impulsado la migración hacia Estados Unidos de América en busca de oportunidades de empleo e ingresos, que no se encuentran en México. Esta migración cobró mayor fuerza en los últimos lustros con la implementación de la apertura comercial y, paradójicamente, coincide con el periodo de mayor recepción de remesas.

Las remesas que vienen del exterior especialmente de Estados Unidos de América, han aligerado los problemas de pobreza, reduciendo ligeramente la cantidad y proporción de hogares en pobreza, tanto en el sector

rural como en el sector urbano. Finalmente, de lo expuesto se desprende que es necesario impulsar el crecimiento y desarrollo económico de México y del sector rural, con políticas de desarrollo intersectorial y regional, de desarrollo agrario, de investigación e innovación tecnológica, entre las más importantes.

REFERENCIAS

- [1] Banco Mundial. (2016). México. Datos de libre acceso del Banco Mundial. [En línea]. Disponible en <<http://datos.bancomundial.org/pais/mexico?view=chart>>
- [2] Caamal C., I. et al. (2014). Pobreza, crecimiento económico, desempleo y salarios en México. En García R., J. F. et al. (eds), Pobreza y desigualdad. Una mirada desde la economía y la investigación operacional (pp. 129-149). RIDECA.
- [3] Caamal C., I. y Pat F., V. G. (2015). Pobreza, migración y remesas en México. En Perales S., A. (coord.), Estudios sobre migración en Norte y Centro América. Chapingo. Universidad Autónoma Chapingo. México
- [4] Consejo Nacional de Población (CONAPO). (2016). Migración. Glosario [En línea] Consultado el 27 de agosto de 2016. México. Disponible en: <http://www.conapo.gob.mx/en/CONAPO/Glosario_Migracion_Interna?page=1>
- [5] Consejo Nacional de Población (CONAPO). (2016). Proyecciones de la población de México. Datos Abiertos. México. [En línea] Disponible en: <<http://datos.gob.mx/busca/dataset/proyecciones-de-la-poblacion-de-mexico>>
- [6] INEGI. (2016). Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2000-2014. [En línea]. México. Disponible en <<http://www.inegi.org.mx/est/contenidos/Proyectos/encuestas/hogares/regulares/enigh/>>.
- [7] INEGI. 2014. Sistema de Cuentas Nacionales. [En línea]. Disponible en <<http://www.inegi.org.mx/est/contenidos/proyectos/scn/>>
- [8] Passel J., et al. (2012). Net Migration from Mexico Falls to Zero—and Perhaps Less. Pew Hispanic Center: Washington, D.C.
- [9] Sánchez Juárez, Leobardo. (2009). Teorías del crecimiento económico y divergencia regional en México. Entelequia, Revista Interdisciplinaria No 9. Primavera. Pp. 129-149.
- [10] Secretaría del Trabajo y Previsión Social (STPS). (2016). Encuesta sobre Migración en la Frontera Norte de México (EMIF-NORTE). México. [En línea] Disponible en: <http://www.stps.gob.mx/bp/secciones/conoce/areas_atencion/areas_atencion/web/menu_infsector.html>
- [11] Statista (2016). Jobs held by people of Mexican origin in the U.S. by economic sector in 2010. [En línea] Consultado el 04 de Julio de 2016. Disponible en: <<http://www.statista.com/search/?statistics=1&forecasts=1&studies=1&industryReports=1&dossiers=1&infos=1&interval=0&category=0&subCategory=0®ion=0&archive=0&q=mexico+immigrants&sortMethod=idrelevance&accuracy=and&itemsPerPage=50&subCategory=0&subCategory=0>>
- [12] Zermeño, F. (2004). Lecciones de desarrollo económico. Ed. Plaza y Valdés, México.

Capítulo 5

pp 51-63

MODELO DE PREDICCIÓN DE SERIES TEMPORALES PARA LA PRODUCCIÓN DE HUEVOS EN LA EMPRESA COMBINADO AVÍCOLA DE PINAR DEL RÍO.

Carlos Jesús Cruz Bracho*, María Amparo León Sánchez*, Eugenio Hernández Rodríguez**
&

Víctor Ernesto Pérez León*

*Universidad de Pinar del Río, Cuba.

**Empresa Avícola de Pinar del Río, Cuba.

ABSTRACT

The present study is an application of time series analysis for predicting the performance of chicken eggs production in the Poultry Company of Pinar del Río, Cuba. The data used includes 146 months from January, 2003 to February, 2015 and the model was conducted through the components decomposition method. Additionally, the behavior of the series "average laying hens" and "efficiency" was analyzed, because of their usefulness in the results interpretation.

Due to the high stationary of the "efficiency" in both climatic periods in Cuba (dry and rainy), the incidence of two climatic variables was studied; furthermore, the variables "age of laying hens" and "food portion" were incorporated. This analysis was conducted by means of a multiple regression equation using successive steps. The predictive model was used to estimate the production from March to December in 2015 and the difference with the real production was inferior to the 0,4%.

KEYWORDS: Time series analysis; multiple regression; egg production

RESUMEN

El presente estudio es una aplicación del análisis de series temporales para predecir el comportamiento de la producción de huevos en la Empresa Combinado Avícola de Pinar del Río, Cuba. Los datos usados abarcan 146 meses, desde enero 2003 hasta febrero de 2015 y el modelo se obtuvo por el método de descomposición de la serie. Se analizaron además, las series "promedio de ponedoras" y "eficiencia" por su utilidad en la interpretación de los resultados.

Debido a la marcada estacionalidad de la eficiencia en ambos períodos climáticos de Cuba (seco y lluvioso) se estudió la incidencia de dos variables climáticas, incluyéndose además la edad promedio del rebaño y la ración promedio diaria de alimento. El análisis se realizó mediante una ecuación de regresión múltiple obtenida por el método de pasos sucesivos. El modelo de pronóstico fue utilizado para estimar la producción de marzo a diciembre de 2015 y la diferencia con la producción real fue inferior al 0,4%

1. INTRODUCCIÓN

En la actualidad, la técnica estadística de análisis de series temporales ha sido ampliamente diseminada en la literatura y existe una gran variedad de circunstancias e investigaciones en que puede ser empleada, especialmente en estudios concernientes a decisiones económicas como el desempeño de economías nacionales e internacionales, la bolsa de valores, la predicción de precios, el suministro y demanda de productos agrícolas, entre otros (de Oliveira et al., 2012).

El presente estudio tiene como objetivo predecir el comportamiento de la producción de huevos de la Empresa Avícola de Pinar del Río (PICAN), que tiene, entre sus principales metas, producir y comercializar aves vivas, huevos de aves, carne de aves, subproductos avícolas. Para ello, se emplearon las series temporales, por su facilidad para, realizar predicciones dado el conocimiento de una variable a través del tiempo.

Existen publicaciones de modelos de producción de huevos que emplean modelos matemáticos y estocásticos. Estos estudios, sin embargo, resultan muy complejos, como el de Grossman & Koops (2001) o requieren demasiadas variables para determinar la producción, como en el caso del modelo estocástico de Alvarez & Hocking (2007), restándoles utilidad práctica en la mayoría de las situaciones comerciales (Ahmand, 2012). Otros modelos, como el modelo no lineal de Gompertz, producen resultados carentes de sentido y no son apropiados para el tipo de variables y datos de producción que emplean (Ahmand, 2012). Existen, además, otras aplicaciones a la producción de huevos como Prasad & Singh (2008), que analiza

varios modelos para describir la producción anual de huevos de gallina, Sumner et al. (2011), Ospina-Álvarez (2013), entre otros.

En ese sentido, el empleo de las series de tiempo en el actual estudio permite a la administración de la entidad obtener mejores resultados, si en alguna medida, se logra reducir la incertidumbre sobre los sucesos futuros. El análisis incluye la estimación de un modelo de pronóstico para la producción de huevos y una comparación del comportamiento de esta serie con las series “promedio de ponedoras” y “eficiencia”.

Para analizar la serie se empleó el Método de Descomposición (Reyes, 2007), que consiste en separar las componentes de una serie: Tendencia, Estacionalidad, Ciclo y Error; este último, constituye el factor aleatorio de la serie. En el estudio no se analiza cada factor para todas las series, sino que se brinda atención a aquellos que proporcionen la información más útil para el análisis de las mismas.

Los resultados reales mostraron una alta confiabilidad para la estimación realizada, pues la diferencia entre la cantidad total estimada, para los últimos ocho meses de 2015 y la real obtenida, no supera el 0,04%.

2. MODELO DE PRONÓSTICO PARA LA PRODUCCIÓN DE HUEVOS

La serie incluye los datos de producción de huevos durante 146 meses, en el período comprendido entre enero de 2003 y febrero de 2015 (en miles de unidades). La producción mínima fue de 6008,10 y la máxima de 23903,10 con una media mensual de 17948,04 miles de huevos.

El comportamiento cronológico que ha tenido la producción en el período que se analiza (Figura1) tiene una depresión significativa en septiembre de 2008, causada por fenómenos meteorológicos (Huracanes Gustav e Ike). Después de la fecha señalada se muestra una recuperación considerable, que se ha mantenido con tendencia creciente, pero menor que la obtenida hasta aquel momento; esto es, la existencia de crecimiento de los niveles de producción de forma sostenida, pero siempre por debajo del último valor alcanzado antes de la depresión de septiembre de 2008.

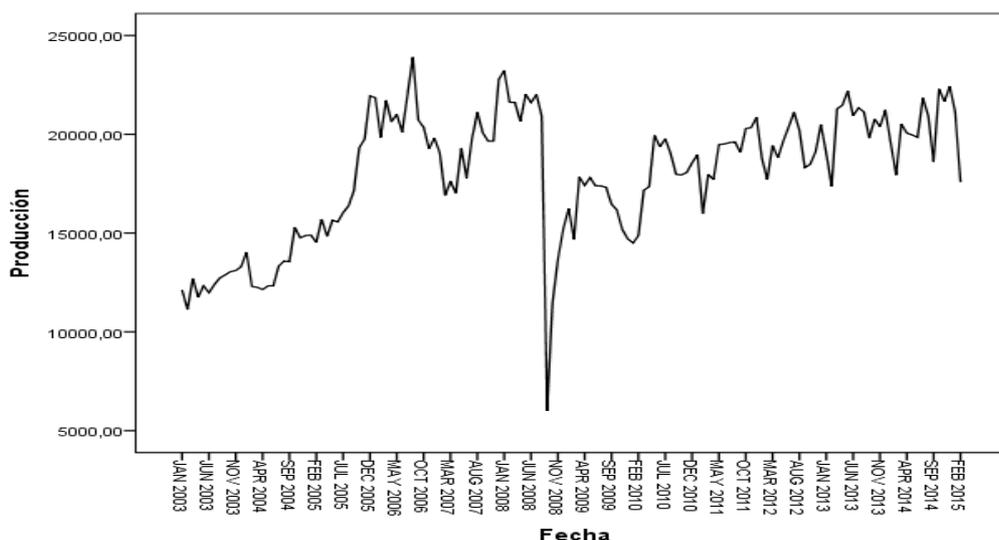


Figura 1. Serie temporal de producción mensual de huevos (Enero 2003-Febrero 2015)

Fuente: Elaboración propia, salida SPSS

Con la finalidad de identificar el modelo que mejor describe el comportamiento de los datos se realizaron varios ajustes, incluyendo toda la información del período analizado (Tabla 1). Todos resultaron significativos, debido a la cantidad de observaciones incluidas; sin embargo, su capacidad predictiva resultó baja (R^2), debido al cambio brusco en la tendencia registrado a partir de septiembre de 2008.

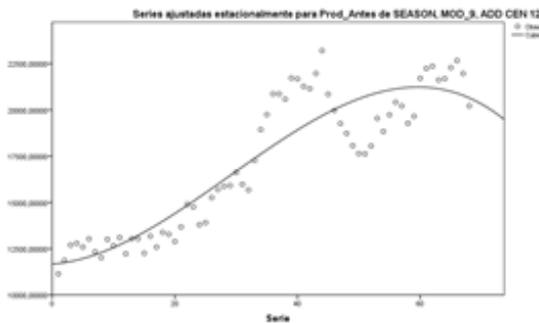
Se dividió la serie en dos etapas: La primera, hasta Septiembre de 2008 (68 datos) y la segunda, desde octubre de 2008 hasta el final del período analizado (78).

Tabla 1. Resumen del modelo y estimaciones de los parámetros

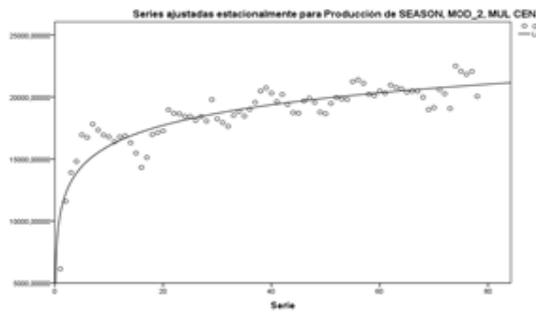
Ecuación	Resumen del modelo					Estimaciones de los parámetros			
	R ²	F	gl1	gl2	Sig.	Constante	b1	b2	b3
Lineal	,347	76,397	1	144	,000	14673,120	44,628		
Logarítmica	,443	114,601	1	144	,000	8893,131	2261,102		
Inversa	,181	31,761	1	144	,000	18476,995	-13741,136		
Cuadrático	,400	47,579	2	143	,000	12995,598	112,636	-,463	
Cúbico	,524	52,017	3	142	,000	9894,216	361,572	-4,682	,019
Compuesto	,332	71,520	1	144	,000	14383,287	1,003		
Potencia	,431	109,212	1	144	,000	10009,336	,141		
S	,182	31,935	1	144	,000	9,810	-,872		
Crecimiento	,332	71,520	1	144	,000	9,574	,003		
Exponencial	,332	71,520	1	144	,000	14383,287	,003		

Fuente: Elaboración propia, salida SPSS

Se ajustaron los modelos para cada una de las etapas establecidas. Ambos resultaron significativos y con alto poder explicativo. El modelo que mejor se ajusta al comportamiento de los datos de la primera etapa es el Cúbico ($R^2 = 0.839$), mientras que el Logarítmico es el que mejor describe la segunda ($R^2 = 0.792$).



Primera Etapa (Hasta Septiembre de 2008)



Segunda Etapa (Después de Septiembre de 2008)

La existencia de un cambio estructural en el comportamiento de los datos es una de las premisas en el pronóstico de series temporales, pues parte del supuesto de que el comportamiento futuro de la serie es similar al de la etapa de análisis. Para demostrarlo en el conjunto de datos analizados se aplicó el test de Chow (Novales, 1996). Este consiste en ajustar el modelo con la muestra total, $n = n_1 + n_2$ denominado modelo restringido, y ajustar los modelos con las muestras separadas, en este caso con $n_1 = 68$, correspondiente al período antes de septiembre de 2008 y $n_2 = 78$, que representa el período después de septiembre de 2008.

Contraste de cambio estructural: Test de Chow

H_0 : Ausencia de cambio estructural

H_1 : Presencia de cambio estructural

$$\frac{[SSR - (SR_1 + SR_2)]/k}{(SR_1 + SR_2)/(n - 2k)} \sim F_{k, n - 2k}$$

donde:

K : cantidad de variables explicativas

SSR : Suma de cuadrados residual del modelo con $n = n_1 + n_2$

SR_1 : Suma de cuadrados residual del modelo con n_1 observaciones

SR_2 : Suma de cuadrados residual del modelo con n_2 observaciones

El estadístico F calculado alcanzó el valor 335,08 muy superior al valor tabulado

$F_{0,95; 2; 142} = 0,36$ por lo que se rechaza la hipótesis nula de ausencia de cambio estructural.

Esta conclusión se corrobora con la figura 1, donde se percibe que el ritmo de crecimiento en el primer período es superior al del segundo. Sin embargo, aunque la tendencia actual de la producción de huevos es creciente, para el final de la serie aún no se han alcanzado los niveles iniciales.

Una vez demostrado el cambio en la estructura de la serie, se decidió no tener en cuenta los datos de antes de enero de 2009. Por lo tanto, de aquí en adelante se analizarán los factores a partir de enero de 2009, fecha a partir de la cual se considera la empresa ya recuperada, en términos de producción.

2.1. Análisis de la serie de producción

Primeramente, se analiza la producción, para la que se determina cuál es el esquema más adecuado para representar la serie (aditivo o multiplicativo). Con esa finalidad, se compararon los coeficientes de variación correspondientes a las primeras diferencias $d_t = \Delta Y_t = Y_t - Y_{t-1}$ y a los cocientes de orden 1 $K_t = \frac{Y_t}{Y_{t-1}}$ de la variable producción. El coeficiente de variación para las diferencias resultó igual a 76,65 y el de los cocientes 0,076, por lo que se seleccionó el esquema multiplicativo.

Con respecto a la estacionalidad, de la figura 3 se desprende que en los meses de diciembre, marzo, mayo, junio y julio de cada año el nivel de la producción está por encima de la media de los meses restantes. Durante el verano, de forma natural, se presenta una sobreproducción de huevos. En diciembre, debido a la necesidad de cumplir los compromisos anuales, la causa de la sobreproducción está dada porque, a veces, se dejan más animales en las granjas. En febrero, los niveles están por debajo de la media, por un problema de calendario, pues es el mes corto, lo que sugiere la presencia de un componente estacional.

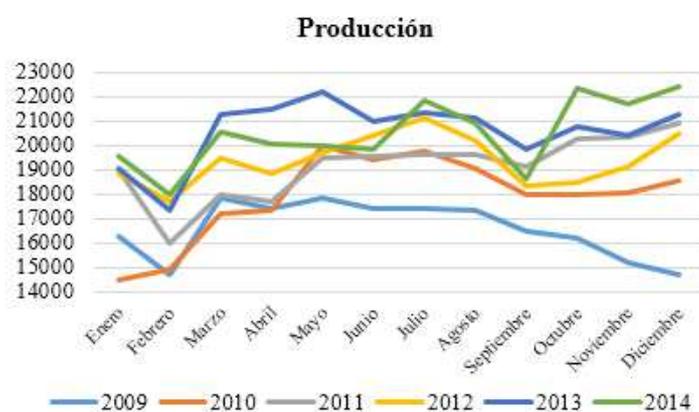


Figura 3. Producción mensual desde 2009 hasta 2014

Fuente: Elaboración propia.

Los coeficientes estacionales, aparecen representados en la figura 4 donde se puede apreciar que la producción tiene sus puntos más bajos cada febrero; sin embargo, no se cuenta con información suficiente para justificar el comportamiento de los meses restantes por lo que se analizaron otras series para buscar respuesta.



Figura 4. Índices de variación estacional de la serie Producción
Fuente: Elaboración propia, salida SPSS

La tendencia de la serie Producción se analiza empleando las medias móviles de orden 12 (Davidson & MacKinnon, 2004), pues estas eliminan los efectos de los componentes estacionales e irregular. La serie se alisa y muestra un comportamiento mucho más suave. El efecto del alisado con las medias móviles es muy conveniente porque se logra depurar la tendencia de la serie original (Davidson & MacKinnon, 2004).

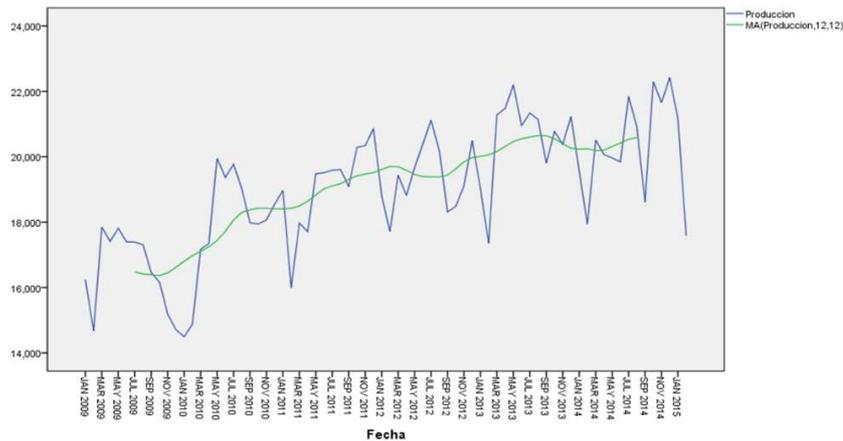


Figura 5. Serie Producción alisada.
Fuente: Elaboración propia, salida SPSS

Como se observa en la figura 5, la tendencia de la Producción (en verde) es a subir, lo que informa que la empresa tiene la capacidad de producir este año más que en años anteriores, aunque el comportamiento de la serie es errático (en azul), para el mes de noviembre de cada año. En ese sentido, al igual que con la estacionalidad, el análisis de la tendencia deberá complementarse con otras series que sirvan de apoyo.

El Componente cíclico de la producción describe cómo, en períodos largos, (mayores al de la estacionalidad) se repiten patrones en la serie.

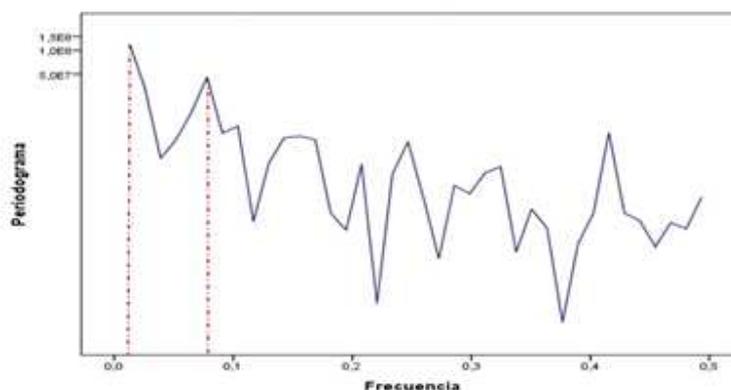


Figura 6. Periodograma de la serie Producción por frecuencia

El pico más alto en el periodograma corresponde a una frecuencia de 0,015 aproximadamente, de ahí que el ciclo de esta serie sea de alrededor de $1/0,015 = 66,7$ meses y el segundo pico más alto, al que corresponde una frecuencia de 0,08, ratifica el período estacional de la serie $1/0,08 = 12$ meses.

2.2 Análisis de las series “eficiencia” y “promedio de ponedoras”

Seguidamente se analizan las series Eficiencia¹ y Promedio de Ponedoras², para apoyar la serie Producción y obtener diferentes pronósticos que ayuden a la empresa a entender mejor cuál es el comportamiento de esta última y porqué. Estas nuevas series a analizar tienen, cada una, 71 valores correspondientes al dato mensual desde enero de 2009 hasta diciembre de 2014.

Tabla 2. Coeficientes de variación

Eficiencia		Promedio de Ponedoras	
$CV_a = 19.76$	$CV_m = 0.0349$	$CV_a = 6.76$	$CV_m = 0.0507$

CV_a : Coeficiente de Variación del modelo Aditivo

CV_m : Coeficiente de Variación del modelo Multiplicativo

Fuente: Elaboración propia

Los coeficientes de variación de ambas series se muestran en la tabla 2. De acuerdo con sus valores, en ambos casos se escoge el modelo multiplicativo.

En los gráficos que se muestran a continuación se analizan la Estacionalidad y la Tendencia de las series Promedio de Ponedoras y Eficiencia (Figuras 6, 7, 8 y 9 respectivamente). Al analizar el componente estacional de esta serie se ve como en diciembre hay un aumento fuera de lo regular en el promedio de animales. Esto se debe a que diciembre es el mes donde se cierra el plan de producción, de modo que, para cumplirlo, se recurre al aumento, casi exponencial, del número de ponedoras. Sin embargo, todavía esta serie no explica por qué en los meses de marzo a junio la producción está por encima de la media; lejos de esto, su factor estacional desfavorece el crecimiento que se evidencia en la producción.

¹ Se calcula dividiendo la cantidad de huevos por gallina promedio de un mes entre los días del mes (%).

² Cantidad promedio diaria de gallinas ponedoras en cada mes.



Figura 6. Índices de variación estacional de la Serie Promedio de Ponedoras
Fuente: Elaboración propia, salida SPSS

La tendencia de esta serie demuestra que el promedio de ponedoras va en aumento, con los años (Figura 7), a pesar de las altas y bajas puntuales que puedan aparecer. Por tanto, se puede concluir que la empresa, en este momento, tiene más animales que nunca antes y, por tanto este factor favorece al aumento que se observa en la producción.

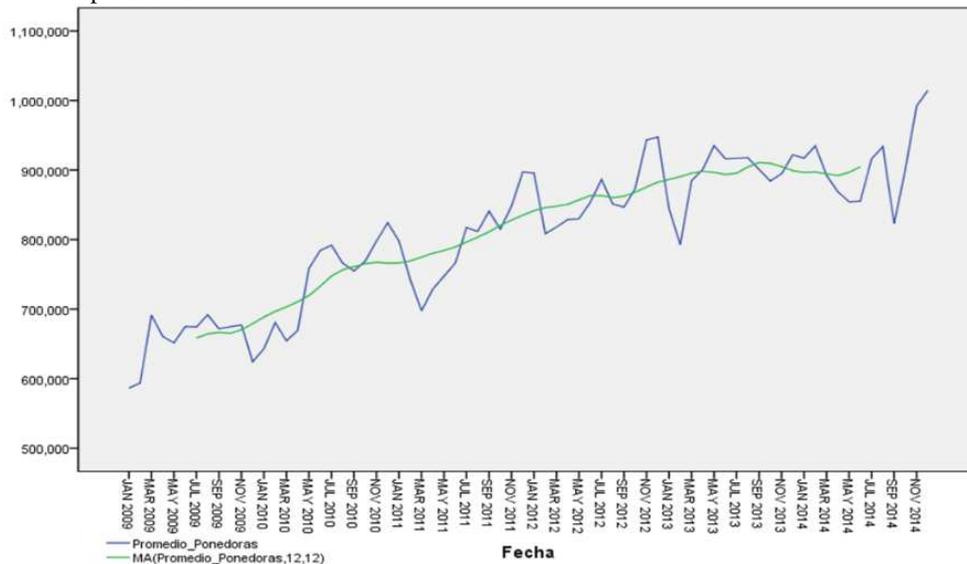


Figura 7. Tendencia en la serie Promedio de Ponedoras
Fuente: Elaboración propia, salida SPSS

La serie Eficiencia es de vital importancia dentro de la empresa y su análisis es de obligada referencia para comprender el comportamiento de la producción. Este indicador, a diferencia de la producción, no depende de los días que tenga el mes, lo que garantiza mayor estabilidad en el tiempo y refleja mejor aquellos factores que lo afectan y de los que depende.

El componente Estacional de la serie Eficiencia es uno de los más claros analizados hasta el momento (Figura 8), ya que muestra cómo durante los meses de verano la eficiencia es alta, mientras que en invierno tiende a estar por debajo de la media. Esto permite asegurar que uno o varios factores climáticos pueden influir en la eficiencia, mientras que factores no climáticos pueden ser los causantes de la inestabilidad de la serie vista en la Figura 9, correspondiente a su tendencia.

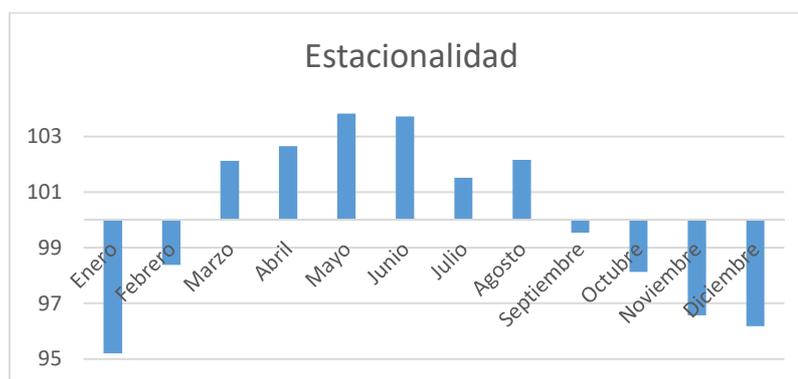


Figura 8. Índices de variación estacional de la serie Eficiencia
Fuente: Elaboración propia, salida SPSS

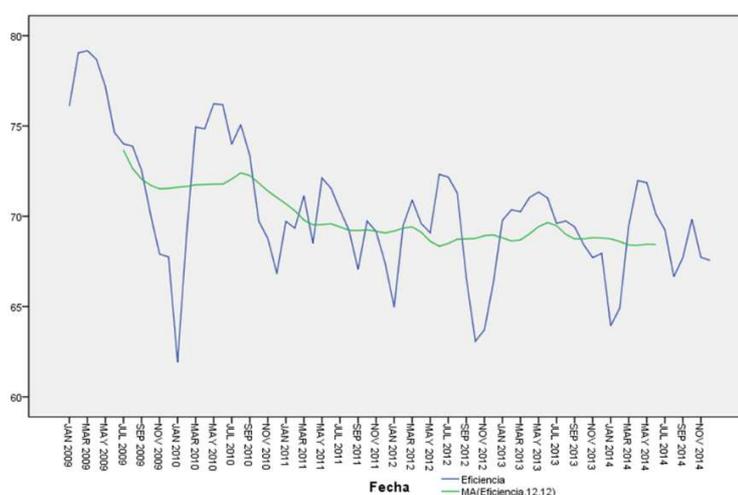


Figura 9. Tendencia de la serie Eficiencia
Fuente: Elaboración propia, salida SPSS

Gráficamente se aprecia que la tendencia de esta serie es a disminuir con el tiempo, lo que lleva a concluir que la empresa tiene problemas con este indicador. De este modo, se puede afirmar que el aumento de la producción marcado anteriormente es debido al incremento del promedio de ponedoras, como ya se indicó.

2.3. Estudio de la influencia de algunas variables sobre la eficiencia.

En esta sección se analiza el impacto que tienen en la Eficiencia los factores “edad promedio” (la edad, en meses, promedio de los animales en la granja), la “ración promedio diaria”, en gramos de alimento, que reciben, la “temperatura”, en grados Celsius y la “humedad relativa” en por ciento.

Para ello se toman los datos de la granja con mayor estabilidad en sus niveles de producción: Granja “El Rosario”, situada en el Municipio Viñales, al Norte de la Provincia de Pinar del Río. La información comprende los meses desde enero de 2004 hasta diciembre de 2014. En este período la eficiencia no presenta un cambio de estructura y, como la producción es pertinente, entonces es posible trabajar con todos estos datos para obtener mayor exactitud en el resultado.

Los indicadores de temperatura y humedad se obtuvieron de la estación más cercana a la granja, la estación de Santa Lucia, en el municipio Minas de Matahambre, ya que la región donde se encuentra la granja carece de estación meteorológica.

Con este análisis se pretende determinar cuáles factores impactan en la eficiencia y entender por qué ha disminuido con los años. Aunque todas las granjas no tienen las mismas características, el alimento es el mismo y la temperatura y humedad son similares por estar todas en la provincia Pinar del Río; por lo tanto, relativamente cerca. Por ello el estudio se realiza en solo una de ellas y se generaliza al resto. Para ello, se

utiliza el método de pasos sucesivos (Weisberg, 2005). El orden de entrada de las variables independientes fue edad, humedad y temperatura (tabla 3)

Tabla 3. Resumen de los modelos
Variables introducidas/eliminadas^a

Modelo	Variables introducidas	Variables eliminadas
1	Edad	
2	Humedad	
3	Temperatura	

a. Variable dependiente: Eficiencia

Fuente: Elaboración propia

La variable ración queda excluida del modelo, lo que lleva a pensar que cantidad de alimento que se brinde a los animales no es un factor determinante, mientras no se violen las especificaciones establecidas. Posiblemente la calidad del alimento sea un factor importante; pero la empresa no cuenta con la información de este indicador a lo largo del tiempo. El ajuste del modelo no es bueno el coeficiente de determinación es bajo y el error típico es superior al 10% del valor medio de eficiencia. (Tabla 4).

Tabla 4. Indicadores de calidad de los Modelos

Modelo	R	R cuadrado	R cuadrado corregida	Error típico. de la estimación
1	,324 ^a	,105	,098	8,21400
2	,378 ^b	,143	,129	8,07058
3	,460 ^c	,211	,193	7,77279

a. Variables predictoras: (Constante), Edad

b. Variables predictoras: (Constante), Edad, Humedad

Fuente: Elaboración propia

Los coeficientes beta tipificados (Tabla 5) dan la humedad como la variable de mayor influencia sobre la eficiencia seguida de la temperatura y la edad. La edad y la humedad, inciden negativamente sobre la eficiencia. La temperatura, por otra parte, es la única con coeficiente positivo, lo que demuestra que valores elevados de temperatura serán favorables; sin embargo, su elevado valor del error hace pensar en que este es un indicador poco fiable. En ese sentido, las condiciones idóneas para alcanzar una eficiencia elevada son animales jóvenes y un clima cálido y seco.

Esto permite concluir que el único factor que se puede controlar es la edad. De este modo, se puede afirmar que los problemas que está presentando la empresa con la eficiencia pueden atenuarse con un adecuado manejo del rebaño. Las gallinas al envejecer se hacen menos eficientes.

Tabla 5. Coeficientes y significación de los modelos estimados

Modelo		Coeficientes no estandarizados		Coeficientes tipificados	t	Sig.
		B	Error típ.	Beta		
1	(Constante)	80,433	1,903		42,275	,000
	Edad	-,899	,231	-,324	-3,895	,000
2	(Constante)	108,878	12,137		8,970	,000
	Edad	-,849	,228	-,307	-3,730	,000
	Humedad	-,375	,158	-,195	-2,372	,019
3	(Constante)	107,115	11,702		9,154	,000
	Edad	-,869	,219	-,314	-3,964	,000
	Humedad	-,797	,199	-,414	-4,016	,000
	Temperatura	1,383	,417	,342	3,316	,001

a. Variable dependiente: Eficiencia

Fuente: Elaboración propia, salida SPSS

A continuación se analizan los factores para las tres series, al mismo tiempo. Para ello, dadas las diferencias de escala, es necesario centrarlas en su valor mínimo. De este modo, quedan expresadas en valores [0,1] que indican los valores máximo y mínimo para cada una.

3. ESTUDIO COMPARATIVO DE LAS SERIES PRODUCCIÓN, EFICIENCIA Y PROMEDIO DE PONEDORAS

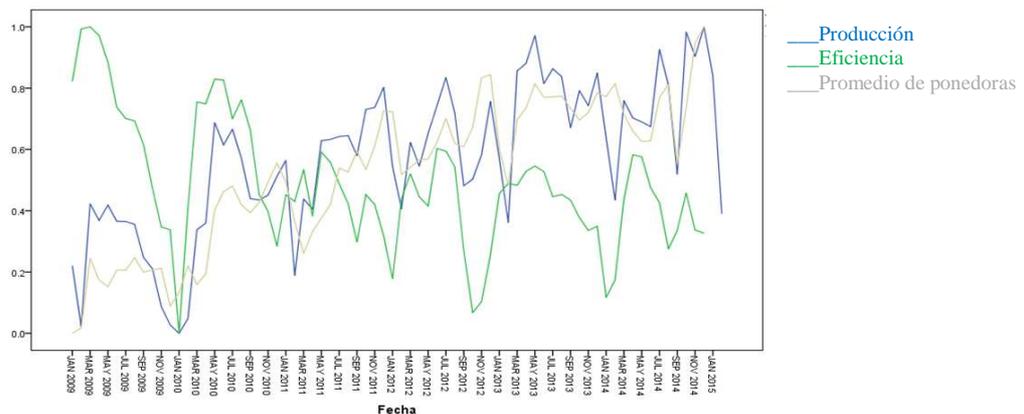


Figura 10. Series Producción, Eficiencia y Promedio de ponedoras centradas en el mínimo
Fuente: Elaboración propia, salida SPSS

La irregularidad de las series en la figura 10 dificulta identificar su comportamiento. Entre ellos, el de mayor variabilidad es el de la eficiencia, que tuvo valores para el año 2009 que no se han alcanzado de nuevo. Con la producción y el promedio pasa lo contrario.

El análisis de las tendencias para estas (Figura 11) demuestra la disminución gradual de la eficiencia durante el período analizado, mientras que el “Promedio de ponedoras” lleva el peso de la producción. Esto indica que se está apostando por la cantidad en vez de por la calidad de los animales. Esta política es errónea porque más animales implica un mayor gasto (indicador no contemplado en el estudio), mientras que trabajar en aras de aumentar la eficiencia por animal, aunque es más difícil, permitiría elevar los niveles de producción. Así mismo, es recomendable realizar un estudio para determinar qué otros factores inciden en la disminución de la eficiencia año tras año, a partir de 2009. Resolver esta incógnita sería de vital importancia en la empresa.

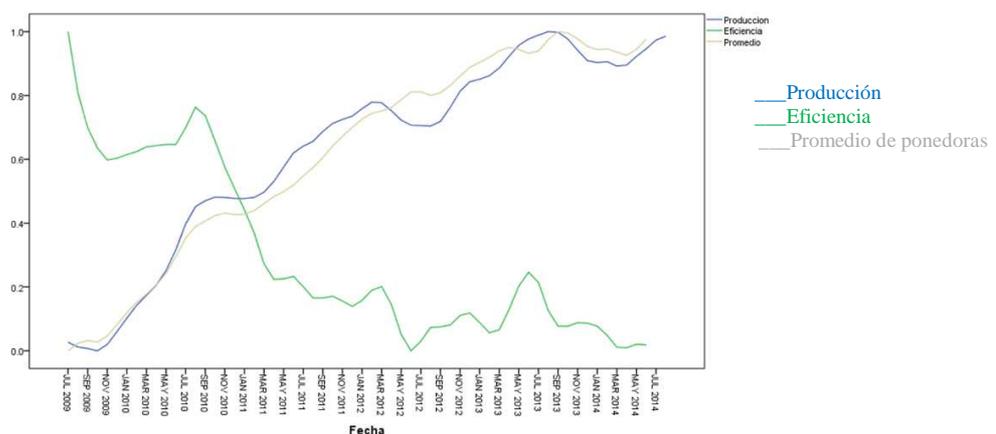


Figura 10. Tendencias de las series Producción, Eficiencia y Promedio de ponedoras centradas en el mínimo
Fuente: Elaboración propia, salida SPSS

Estacionalidad



Figura 11. Índices de variación estacional en todas las series
Fuente: Elaboración propia

La figura 11 constituye uno de los gráficos con mayor poder explicativo del estudio. Este permite apreciar la influencia conjunta de los tres factores estacionales. En febrero, valores desfavorables para la eficiencia y el promedio de ponedoras provocan un gran descenso en la producción. Por su parte, en diciembre se evidencia que el aumento del rebaño trae consigo un aumento también del nivel de producción, pero disminuye la eficiencia. Lo que corrobora los resultados previos obtenidos.

A partir de aquí, con todos los factores analizados, es conveniente realizar la predicción, pues ya se ha descompuesto la serie Producción y se procede a elegir el modelo de regresión que más se ajuste a la serie desestacionalizada, ya que omitir el factor estacional dará más exactitud a la estimación de la tendencia.

Los resultados aparecen en la tabla 6, donde se muestra que el modelo cúbico y el cuadrático son los de mayor ajuste, siendo el mejor, el cúbico. Por tanto, la ecuación de regresión queda de la siguiente forma:

$$\hat{Y}_t = 16\,219,829 + 79,375X + 0,439X^2 - 0,09X^3$$

Tabla 6. Resumen del modelo y estimaciones de los parámetros de la serie desestacionalizada

Variable dependiente: Series ajustadas estacionalmente para Producción

Ecuación	Resumen del modelo					Estimaciones de los parámetros			
	R ²	F	gl1	gl2	Sig.	Constante	b1	b2	b3
Lineal	,692	161,842	1	72	,000	16592,842	65,413		
Logarítmica	,594	105,398	1	72	,000	14269,177	1427,746		
Inversa	,178	15,636	1	72	,000	19398,197	-5334,719		
Cuadrático	,714	88,572	2	71	,000	16015,743	110,973	-,607	
Cúbico	,716	58,730	3	70	,000	16219,829	79,375	,439	-,009
Compuesto	,678	151,292	1	72	,000	16635,175	1,004		
Potencia	,591	103,825	1	72	,000	14661,334	,077		
S	,178	15,568	1	72	,000	9,870	-,288		
Crecimiento	,678	151,292	1	72	,000	9,719	,004		
Exponencial	,678	151,292	1	72	,000	16635,175	,004		

Fuente: Elaboración propia, salida SPSS

Una vez hallado el modelo de regresión que más se ajusta a la serie y los índices de variación estacional es posible predecir los valores que se deseen. Se realizó una predicción hasta diciembre de 2015 con el fin de comprobar si el plan de la empresa de 249.2 millones de huevos es posible. La predicción parece bastante buena a simple vista (figura 12), cuando se compara con la producción real, en muchos valores da una medida de que la estimación es acertada.

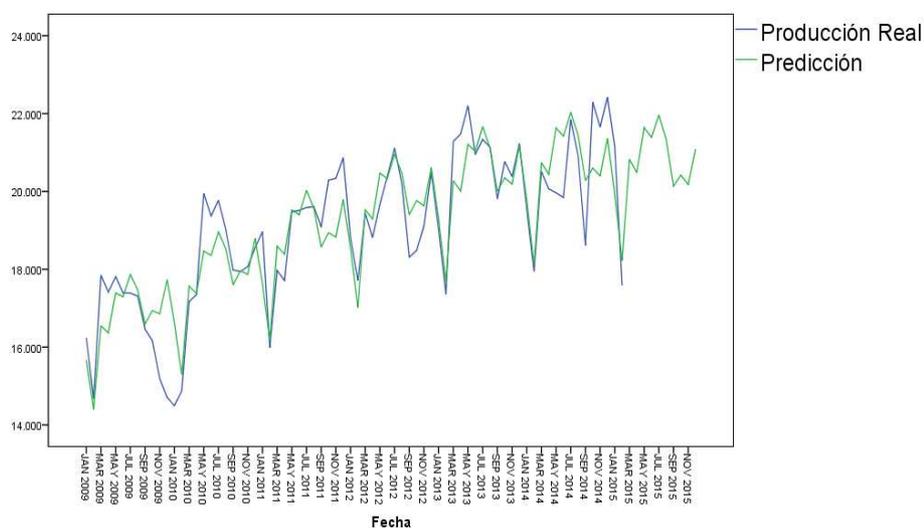


Figura 12. Producción real y estimada hasta diciembre de 2015

Fuente: Elaboración propia, salida SPSS

4. CONCLUSIONES

El presente estudio es una aplicación del análisis de Series Temporales para la predicción de la producción de huevos en la Empresa Avícola de Pinar del Río, Cuba. La investigación incluyó información correspondiente a 146 meses, desde 2003 hasta 2015.

Se pudo constatar que el Promedio de ponedoras y la eficiencia de estas inciden directamente en la producción, lo cual es un aspecto controlable. Por otra parte, la edad de las ponedoras, la temperatura y la humedad tienen incidencia en la eficiencia, no así la cantidad de alimento, el que se puede afirmar que es satisfactorio, a pesar de que se desconoce la calidad de su constitución.

En cuanto a los niveles de producción y eficiencia, el primero no ha logrado valores obtenidos de antes de 2009 y la eficiencia ha bajado con el paso de los años, mientras que ha aumentado el promedio de ponedoras, que es el causante del aumento de la producción actualmente. Con respecto a este último, existe un factor estacional que favorece los niveles de producción en marzo y agosto, y los dificulta de septiembre a febrero.

Hasta el momento no ha sido posible recuperar los niveles de producción de antes de 2008. La estacionalidad que muestra un aumento de producción en el verano no va acompañada de un efectivo tratamiento del rebaño en esos meses. La reserva factible para garantizar una eficiencia adecuada es el manejo oportuno del rebaño. La eficiencia ha bajado desde 2009 y no tiende a remontar.

El promedio de ponedoras es el pilar en que la empresa está basando el cumplimiento de los planes de producción.

Futuros estudios deben estar encaminados a analizar la calidad del pienso para así conocer su impacto en la eficiencia.

REFERENCIAS

- [1] Ahmad H. A. (2012): Egg production forecasting: Determining efficient modeling approaches. Journal of Applied Poultry Research 20(4):, 463-473.

- [2] Alvarez R., & Hocking P. M. (2007): Stochastic model of egg production in broiler breeders. *Poult Sci* (86):, 1445–1452.
- [3] Davidson, R., & MacKinnon, J. G. (2004):. *Econometric theory and methods* (Vol. 5):. New York: Oxford University Press.
- [4] de Oliveira, S, C., Marcondes, L. M., Sakaguchi J. T., & Carvalho, P. (2012): A study about the performance of time series models for the analysis of agricultural prices. *GEPROS. Gestão da Produção, Operações e Sistemas*, Ano 7, nº 3, jul-set/2012, 11-27.
- [5] Grossman M. & Koops W. J. (2001): A model for individual egg production in chickens. *Poult Sci* (80):, 859–867.
- [6] Ospina-Álvarez A., Bernal, M., Catalán, I. A., Roos, D. Bigot J. L., & Palomera I. (2013): Modeling fish egg production and spatial distribution from acoustic data: A step forward into the analysis of recruitment. *PLoS ONE* 8(9):, 1-19.
- [7] Prasad Shiv & Singh D. P. (2008): Mathematical model for annual egg production curves of chicken. *Indian Journal of poultry Science* 43(3):, 293-295.
- [8] Reyes Aguilar, P. (2007): Metodología de análisis con series de tiempo. <http://www.icicm.com/files/SeriesDeTiempoComp.doc> [Acceso Sept 2015]
- [9] Sumner D. A., Gow H., Hayes D., Matthews W, Norwood B, Rosen-Molina J. T., & Thurman W. (2011): Economic and market issues on the sustainability of egg production in the United States: Analysis of alternative production systems. *Emerging Issues: Social Sustainability of Egg Production Symposium at the joint annual meeting of the Poultry Science Association* pp, 1-10. doi:10.3382/ps.2010-00822

ANÁLISIS DE LA SENSIBILIDAD ALÉRGICA OCUPACIONAL EN PANADERÍAS DE LA HABANA

Sahily Estradé Fernández*, Vivian Sistachs Vega*, Dra. Mirta Álvarez Castelló**

*Universidad de La Habana, ** Hospital "Calixto García"

ABSTRACT

Is of great importance to know about the health situation of bakery workers, since allows the specialists on the field of medicine to be able to implement a diagnostic method to prevent occupational allergy respiratory diseases on this sector. This paper has the goal to detect factors that have an influence on the baker's sensitivity towards allergens present on their working environment. In the development of this work we use multivariate statistics divided in two phases, the first an observational study from the application of Principal Component Analysis and the second phase a confirmatory study from a Logistic Regression model applied to a Case-Control design.

It was proved that the allergens from the environment and from the working space on the bakeries are the main factor behind the incidence of allergic sensitivity on bakers and other individuals studied.

KEYWORDS: Logistic Regression, Principal Components Analysis, Case-Control studies, Agents.

RESUMEN

Es de gran importancia conocer sobre la situación de salud de los trabajadores de las panaderías, ya que esto permite a los especialistas en el campo de la medicina implementar un método de diagnóstico para la prevención de enfermedades respiratorias alérgicas ocupacionales en este sector. En este artículo se planteó como objetivo detectar factores que influyen en la sensibilidad de los panaderos hacia alérgenos presentes en este sector laboral. En el desarrollo del mismo, se hizo uso de técnicas estadísticas multivariadas divididas en dos fases, la primera un estudio observacional a partir de la aplicación de Análisis de Componentes Principales y la segunda, un estudio confirmatorio partir de un modelo de Regresión Logística aplicado a un diseño Caso-Control. Demostrándose que los alérgenos tanto del medio ambiente como del sector de las panaderías son el factor fundamental que está incidiendo en la sensibilización alérgica presente en los panaderos y en los restantes individuos estudiados.

1. INTRODUCCIÓN

La relación entre salud y condiciones de trabajo es tan antigua como el hombre mismo. En la actualidad es una responsabilidad de la salud pública no solo la vigilancia de la salud del trabajador, también el control y tratamiento de las enfermedades crónicas, entre estas las alérgicas. La prevalencia de estas enfermedades se ha incrementado de manera alarmante en las últimas décadas y ha aumentado su complejidad y severidad, acompañadas de elevada morbimortalidad, generando gastos excesivos para los pacientes y el estado.

La actividad de los panaderos una de las profesiones más vinculadas al desarrollo del asma ocupacional, reportándose una fuerte asociación entre la exposición al polvo de harina y la sensibilización a los alérgenos. Otra exposición de importancia en estos trabajadores es a los ácaros presentes en el polvo o que contaminan los sacos de harina. Representando un factor causal de sensibilización alérgica, que es relevante no sólo por su asociación con enfermedades alérgicas, sino también por el alto porcentaje de la población que está sensibilizado a ellos.

En Cuba no se conoce exactamente la prevalencia de sensibilización al trigo en la población general, ni en trabajadores expuestos, aunque constituye un alérgeno de importancia ocupacional pues está bien identificado como una de las causas de asma de origen laboral.

Es por ello la necesidad de llevar a cabo investigaciones de este tipo entre los sujetos expuestos a estos alérgenos para estudiar la sensibilidad a alérgenos presentes en el entorno ocupacional de este sector laboral en Cuba, incluyendo la sensibilización a especies de ácaros endémicas de países tropicales.

Por eso los objetivos que se persiguen este artículo son describir la relación entre los diferentes alérgenos presentes en la actividad laboral de los panaderos, analizar factores de riesgo alérgico presentes

en el ambiente de panaderías de La Habana y estudiar el comportamiento de tales factores a partir de modelo de regresión logística para un diseño de caso-control y ayudar a resolver la situación de los panaderos de La Habana.

Para el logro de lo anterior; en el epígrafe dos se trata la fundamentación teórica, tanto del Análisis de Componentes Principales como los aspectos generales de la Regresión Logística y en particular el diseño Caso-Control. Por último, en el epígrafe 3 se discuten los resultados.

2. FUNDAMENTACIÓN TEÓRICA

En esta sección se expondrá de forma breve los métodos utilizados para la realización de la descripción de los datos, como son el Análisis de Componentes Principales, la Regresión Logística Binaria y qué es un estudio Caso-Control.

2.1. Análisis de Componentes Principales (ACP)

La idea fundamental del método conocido como Análisis de Componentes Principales (ACP) consiste en reducir la dimensión de los datos derivados de la medición a un grupo de individuos u objetos de un número elevado de variables cuantitativas interrelacionadas entre sí. Esta técnica constituye una transformación lineal que escoge un nuevo sistema de coordenadas para el conjunto original de datos en el cual la varianza de mayor tamaño del conjunto de los datos es capturada en el primer eje (llamado primer componente principal), la segunda varianza mayor es el segundo eje y así sucesivamente. (ver [3])

En el ACP, la información inicial es una tabla bidimensional, (Tabla 1) que cruza n individuos w_i y m variables cuantitativas x_k . Generalmente, las filas de la tabla representan los individuos y las columnas las variables. La posición (i, k) de la tabla se identifica con el valor que representa la variable k con el individuo i .

		Variables			
		X_1	X_2	...	X_m
Individuos	w_1				
	w_2				
	\vdots				
	w_N				

Tabla 1 Información inicial

El ACP permite encontrar un número de factores subyacentes $p < m$ que explican aproximadamente el valor de las m variables para cada individuo. El hecho de que existan estos p factores puede interpretarse como una reducción de la dimensionalidad de los datos: donde antes necesitábamos m valores para caracterizar a cada individuo ahora nos bastan p valores. A cada uno de los p se les llama componente principal. (ver [3][5])

2.2. Regresión Logística

La Regresión Logística es un modelo de regresión para variables dependientes o de respuesta binomialmente distribuidas. Los modelos de regresión logística binaria resultan los de mayor interés ya que la mayor parte de las circunstancias analizadas en medicina responden a este modelo (presencia o no de enfermedad, éxito o fracaso, etc.). Es útil para modelar la probabilidad de un evento que ocurre como función de otros factores. La variable dependiente será una variable dicotómica Y que se codificará como 0 o 1 (respectivamente, “ausencia” y “presencia”), para una serie de k -variables de regresión $X = (x_1, \dots, x_k)$ que llamaremos variables predictoras o covariables.

Para predecir esta probabilidad se usa frecuentemente la función **logit** o logaritmo de las verosimilitudes. Esta función es la base del cálculo de una cierta probabilidad P que se quiere predecir.

La ecuación de partida en los modelos de regresión logística es:

$$\pi(x_i) = P(Y = 1|X) = \frac{\exp(\beta_0 + \sum \beta_k x_k)}{1 + \exp(\beta_0 + \sum \beta_k x_k)}$$

Nótese que en la expresión anterior, β_0 es una constante y los β_i ($i = 1, \dots, k$) son los coeficientes de las covariables x_i correspondientes. Además de esa expresión, obtenemos el cociente de probabilidades, conocido como odds ratio, (ver [4][6])

$$\frac{\pi(x_i)}{1 - \pi(x_i)} = \exp\left\{\beta_0 + \sum \beta_k x_k\right\}$$

2.3. ¿Qué es un estudio Caso-Control (CC)?

Un estudio caso-control, es un estudio epidemiológico, observacional, analítico, en el que los sujetos son seleccionados en función de que tengan (casos) o no tengan (control) una determinada enfermedad, o en general un determinado efecto. Una vez seleccionados los individuos en cada grupo, se investiga si estuvieron expuestos o no a una característica de interés y se compara la proporción de expuestos en el grupo de casos frente a la del grupo de controles.

2.3.1. Modelo de regresión logística para un diseño de casos y controles

Primero introduciremos la variable Z que describe la presencia o no de un individuo en la muestra (1 = caso, 0 = control).

Además, $q_1 = P(Z = 1|Y = 1)$ denota la probabilidad de que la muestra sea un caso y $q_0 = P(Z = 1|Y = 0)$ denota la probabilidad de que la muestra sea un control.

Necesitamos un modelo para $P(Y = 1|Z = 1, X = x)$, asumiendo que $P(Y = 1|X = x)$ sigue un modelo logístico.

$$P(x) = P(Y = 1|Z = 1, X = x) = \frac{q_1 \exp(\beta_0 + \beta_1 x)}{q_0 + q_1 \exp(\beta_0 + \beta_1 x)}$$

Dividiendo el numerador y el denominador por q_0 y usando $\frac{q_1}{q_0} = \exp\left\{\ln\left(\frac{q_1}{q_0}\right)\right\}$, obtenemos,

$$P(Y = 1|Z = 1, X = x) = \frac{\exp(\beta_0^* + \beta_1 x)}{1 + \exp(\beta_0^* + \beta_1 x)}$$

con $\beta_0^* = \beta_0 + \ln\left(\frac{q_1}{q_0}\right)$
ver [1].

3. RESULTADOS Y DISCUSION

Para procesar los datos se utilizó como asistente de software IBM SPSS Statistics versión 20.

En esta investigación se tomó como muestra 120 individuos trabajadores en panaderías pertenecientes a zonas urbanas de La Habana (Guanabacoa, San Miguel del Padrón, Plaza, Habana Vieja y Centro Habana). Además se contó con otros 120 individuos que llamaremos grupo control, no expuesto al ambiente de panadería, pero que también presentaban problemas de alergias.

3.1. Diseño de la investigación

Se diseñó un estudio en dos fases:

- La primera, consistió en un estudio exploratorio a partir de la aplicación de Análisis de Componentes Principales el cual nos permite describir, de un modo sintético, la estructura y las interrelaciones de las variables originales en el fenómeno que se estudia a partir de las componentes obtenidas.

- La segunda fase consistió en un estudio confirmatorio a partir de un modelo de Regresión Logística aplicado a un diseño Caso-Control, para identificar los factores de riesgos asociados a una mayor sensibilización a los alérgenos testados.

El siguiente cuadro relaciona las variables estudiadas junto a su escala de medida

Cuadro1 Variables de interés

Variabes	Medidas
Sensibilización	Dicotómica
Sexo	Dicotómica
Edad	Dicotómica
Fumar	Dicotómica
DP	Catégoricas
Blomia	Catégoricas
DS	Catégoricas
Siro	Catégoricas
Farinae	Catégoricas
Tyrophagus	Catégoricas
Lepidoglyphus	Catégoricas
Soya en polvo	Catégoricas
Trigo	Catégoricas
Levadura	Catégoricas

3.2. Análisis de los resultados

3.2.1. Primera fase

Al aplicar el Análisis de Principales Componentes a los datos, se puede apreciar en la Tabla 2 la existencia de autovalores mayores que 1, dando el número necesario de componentes para la investigación del comportamiento de los mismos.

Componente	Autovalores iniciales			Sumas de extracción de cargas al cuadrado		
	Total	% de varianza	% acumulado	Total	% de varianza	% acumulado
1	3,588	27,603	27,603	3,588	27,603	27,603
2	2,292	17,629	45,233	2,292	17,629	45,233
3	1,145	8,805	54,038	1,145	8,805	54,038
4	1,002	7,706	61,744			
5	,797	6,134	67,878			
6	,759	5,835	73,713			
7	,743	5,717	79,431			
8	,614	4,723	84,153			
9	,562	4,320	88,474			
10	,482	3,705	92,178			
11	,467	3,592	95,771			
12	,289	2,225	97,995			
13	,261	2,005	100,000			

Método de extracción: análisis de componentes principales.

Tabla 2 Varianza total explicada

Aunque se puede observar que el por ciento acumulado de varianza no es muy bueno (54 %). Es posible, que existan otras variables a incluir con la ayuda del especialista, con lo cual mejoraríamos el por ciento de varianza acumulada. En un estudio anterior, se incluyó los años de exposición dando un mejor por ciento, pero en esta investigación no fue posible añadirla debido a que se desea realizar un estudio caso-control y los controles no presentan esta variable.

	Componente		
	1	2	3
Sexo	,358	,117	,500
Edad	-,197	,322	,234
Fumar	-,035	,409	,681
DP	,711	-,470	,064
Blomia	,693	-,399	-,038
DS	,742	-,439	-,071
Siro	,579	,359	-,117
Farinae	,801	-,204	,081
Tyrophagus	,513	,365	,241
Lepidoglyphus	,568	,283	,142
Soya_en_polvo	,308	,553	-,407
Trigo	,303	,607	-,187
Levadura	,368	,613	-,259

Método de extracción: análisis de componentes principales.

Tabla 3 Matriz de componentes

Al observar la Tabla 3 cuando analizamos los valores de cada una de las componentes, se puede llegar a varias conclusiones,

- La primera componente, muestra valores altos en los alérgenos del medio ambiente con relación a las otras variables tomadas en consideración, dando a entender que estos alérgenos están influyendo en el estado de salud de los individuos en estudio.
- La segunda componente, tiene un comportamiento similar, pero asociado a los alérgenos presentes en las panaderías (soya en polvo, trigo y levadura).
- La tercera componente, nos muestra la existencia de una relación entre el sexo (masculino) y el hábito de fumar.

De este modo, con el análisis de componentes principales se obtuvieron la relación de las tres causas presentes en el problema.

3.2.2. Segunda fase

De acuerdo con la teoría desarrollada anteriormente para la realización de un estudio Caso-Control, se tomó la variable dicotómica Z para representar los casos ($Z = 1$) y los controles ($Z = 0$), además la variable de respuesta Y que representa la presencia ($Y = 1$) o ausencia ($Y = 0$) de sensibilización hacia los alérgenos (DP, Blomia, DS, Siro, Farinae, Tyrophagus, Lepidoglyphus, soya en polvo, trigo y levadura) en cada uno de los 224 individuos.

Al aplicar el modelo de Regresión Logística Binaria para el estudio Caso-Control, se obtuvieron los resultados siguientes,

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	33,106 ^a	,677	,942

a. La estimación ha finalizado en el número de iteración 12 porque las estimaciones de los parámetros han cambiado en menos de ,001.

Tabla 4 Resumen del modelo

En la Tabla 4 se muestra la validez de forma global del modelo.

- La prueba $-2 \log$ de la verosimilitud ($-2LL$) mide hasta qué punto un modelo se ajusta bien a los datos, y hay que tener en cuenta que mientras más pequeño sea $-2LL$, mayor es la verosimilitud y mejor ajustado está el modelo. En este caso tiene un valor de 33.106.

- La prueba R cuadrado de Cox y Snell: es un coeficiente de determinación generalizado que se utiliza para estimar la proporción de varianza de la variable dependiente explicada por las variables predictoras (independientes). La R cuadrado de Cox y Snell se basa en la comparación del *log* de la verosimilitud (*LL*) para el modelo respecto al *log* de la verosimilitud (*LL*) para un modelo de línea base. Sus valores oscilan entre 0 y 1. En nuestro caso su valor es 0.677, el cual indica que el 67.7% de la variación de la variable dependiente es explicada por la variable incluida en el modelo.
- La R cuadrado de Nagelkerke: es una versión corregida de la R cuadrado de Cox y Snell. La R cuadrado de Cox y Snell tiene un valor máximo inferior a 1, incluso para un modelo "perfecto". La R cuadrado de Nagelkerke corrige la escala del estadístico para cubrir el rango completo de 0 a 1. En nuestro caso este toma el valor 0.942.

		Sensibilización = 0		Sensibilización = 1		Total
		Observado	Esperado	Observado	Esperado	
Paso 1	1	24	24,000	0	,000	24
	2	24	23,997	0	,003	24
	3	24	23,225	0	,775	24
	4	6	7,578	18	16,422	24
	5	1	,198	23	23,802	24
	6	0	,003	25	24,997	25
	7	0	,000	24	24,000	24
	8	0	,000	24	24,000	24
	9	0	,000	10	10,000	10
	10	0	,000	37	37,000	37

Tabla 5 Tabla de contingencia para la prueba Hosmer-Lemeshow

Paso	Chi cuadrado	gl	Sig.
1	4,564	8	,803

Tabla 6 Prueba Hosmer-Lemeshow

Pasando a la realización de la prueba de Hosmer-Lemeshow, la cual consiste en observar detalladamente cómo se separan los valores observados respecto a los valores esperados según las predicciones del modelo de regresión ajustado, se puede apreciar en la Tabla 5 que el ajuste es bueno, ya que existen coincidencias entre los observados y los valores esperados (ver [6]). Además al contrastar estos valores con la prueba chi-cuadrado se tiene un valor de esta significativamente pequeño (ver Tabla 6), por lo que existe un buen ajuste del modelo a los datos.

Observado	Sensibilización	Pronosticado		Porcentaje correcto
		Sensibilización		
		0	1	
Paso 1	0	75	4	94,9
	1	2	159	98,8
Porcentaje global				97,5

a. El valor de corte es ,500

Tabla 7 Tabla de clasificación

Siguiendo con el análisis, la Tabla 7 nos proporciona la clasificación de los casos según su ocurrencia y según la predicción realizada del modelo. Como puede observarse el porcentaje total de acierto en el pronóstico de sensibilidad es 97.5%.

Por último, la Tabla 8 ilustra las variables presentes en el modelo.

	B	E.T.	Wald	gl	Sig.	Exp(B)
Paso 1 ^a Sexo	1,219	1,254	,945	1	,331	3,385
Edad	,035	,038	,815	1	,367	1,035
Fumar	-,385	1,148	,112	1	,738	,681
DP	-,062	,366	,029	1	,865	,940
Blomia	2,345	,871	7,251	1	,007	10,435
DS	1,784	,591	9,114	1	,003	5,956
Siro	,829	,541	2,344	1	,126	2,290
Farinae	2,730	1,008	7,333	1	,007	15,327
Tyrophagus	1,308	,431	9,208	1	,002	3,698
Lepidoglyphus	1,564	,524	8,906	1	,003	4,780
Soya_en_polvo	1,090	,518	4,427	1	,035	2,974
Trigo	2,242	,821	7,449	1	,006	9,413
Levadura	,982	,550	3,186	1	,074	2,671
Constante	-16,044	5,234	9,395	1	,002	,000

a. Variable(s) introducida(s) en el paso 1: Sexo, Edad, Fumar, DP, Blomia, DS, Siro, Farinae, Tyrophagus, Lepidoglyphus, Soya_en_polvo, Trigo, Levadura.

Tabla 8 Variables en la ecuación

La siguiente relación son las variables que resultaron significativas, entre paréntesis aparece su significación,

- Blomia (0.007)
- Ds (0.003)
- Farinae (0.007)
- Tyrophagus (0.002)
- Lepidoglyphus (0.003)
- Soya en polvo (0.035)
- Trigo (0.006)
- Levadura (0.074)

Además poseen un valor OR de,

- Blomia (10.435)
- Ds (5.956)
- Farinae (15.327)
- Tyrophagus (3.698)
- Lepidoglyphus (4.780)
- Soya en polvo (2.974)
- Trigo (9.413)
- Trigo (9.413)

Lo que conlleva a suponer que todos son factores de riesgo en la presencia de sensibilización alérgica.

Luego, nos encontramos con casos como son el de la variable sexo y el ácaro Siro que no salen como variables significativas en el modelo (0.331 y 0.126 respectivamente), pero si como factores de riesgo ya que sus OR son 3.385 y 2.290 respectivamente.

Además, siguiendo los pasos de un estudio caso-control explicado anteriormente tenemos,

$$q_1 = P(Z = 1|Y = 1) = 0.51$$

$$q_0 = P(Z = 1|Y = 0) = 0.48$$

y finalmente como valor del término independiente,

$$\beta_0^* = -0.08 + \ln\left(\frac{0.51}{0.48}\right) = -0.02$$

Como se puede apreciar, posee un valor poco significativo para el modelo, por lo que podemos despreciarlo.

Del análisis de la tabla de clasificación (ver Tabla 7) del modelo obtenemos la especificidad (proporción entre frecuencias de aciertos negativos y frecuencia total de negativos observados) y la susceptibilidad (razón entre la frecuencia de aciertos positivos y la frecuencia total de positivos observados).

Por ejemplo:

- Especificidad: 0.949
- Susceptibilidad: 0.988

De los datos anteriores podemos señalar que las variables del modelo tienen alta susceptibilidad para diagnosticar adecuadamente la sensibilidad alérgica, ya que hay un por ciento muy alto en el

diagnóstico (99% aproximadamente), también tiene una alta especificidad (95% aproximadamente), es decir su capacidad de detectar casos que no presentan sensibilidad alérgica es muy bueno.

Una mejor y más completa descripción de una tabla de clasificación es el área bajo la curva ROC.

En el siguiente cuadro, se detalla que es una curva ROC y la interpretación de su área.

La curva ROC, es un gráfico donde se observan todos los pares susceptibilidad-especificidad (S-E) resultantes de la variación continua de los puntos de corte en todo el rango de resultados observados, cada punto de la curva representa un par S-E correspondiente a un nivel de decisión determinado. Una prueba con discriminación perfecta, tiene una curva ROC que pasa por la esquina superior izquierda, donde S y E toman valores máximos, es decir ambos son iguales a uno. (ver [2][4][6])

Área bajo la curva ROC, es una medida global de la exactitud de la prueba diagnóstica. Se define como la probabilidad de clasificar correctamente un par de individuos sano y enfermo, seleccionados al azar de la población mediante los resultados obtenidos al aplicarles la prueba diagnóstica. El área bajo la curva siempre es mayor o igual que 0.5, se interpreta que cuando esta área se encuentra entre 0.5 y 0.7 estas indican baja exactitud, entre 0.7 y 0.9 indican que pueden ser útiles para algunos pronósticos y un valor mayor que 0.9 indican exactitud alta. (ver [2][4][6])

Para nuestro modelo, obtenemos la curva ROC, Figura 1 la cual muestra gráficamente el buen pronóstico que se obtiene del modelo de regresión logística para un diseño caso-control, ya que como se observa tanto la especificidad como la susceptibilidad toman valor uno.

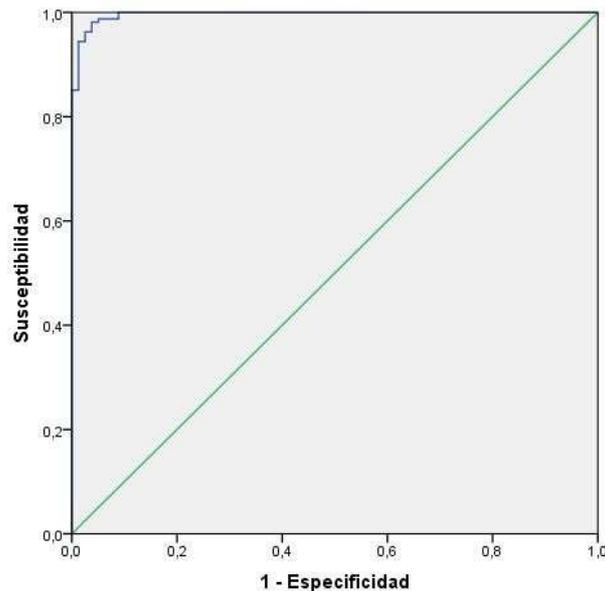


Figura 1 Curva ROC

Ahora, para observar cuantitativamente este buen diagnóstico, se pasa a calcular el área bajo la curva ROC, que como se puede apreciar en la Tabla 9 y por lo comentado anteriormente, nuestro diagnóstico tiene una exactitud alta ya que su área es igual a 0.996.

Variables resultado de contraste: Probabilidad pronosticada

Área	Error típ. ^a	Sig. asintótica ^b	Intervalo de confianza asintótico al 95%	
			Límite inferior	Límite superior
,996	,002	,000	,992	1,000

a. Bajo el supuesto no paramétrico

b. Hipótesis nula: área verdadera = 0,5

Tabla 9 Área bajo la curva

4. CONCLUSIONES

Los resultados obtenidos confirman que los alérgenos tanto del medio ambiente como los propios del sector de las panaderías son el factor fundamental que está incidiendo en la sensibilización alérgica presente en los panaderos y en los restantes individuos estudiados.

Es de suma importancia el conocimiento de estos resultados, ya que de esta forma se podrá ayudar a la preservación de la salud de los trabajadores y a la eliminación de enfermedades causadas por el medio ambiente del trabajo o la propia actividad laboral, evitando pérdidas de días laborales y gastos en medicamentos por enfermedad de los individuos y gastos al país.

REFERENCIAS

- [1] Agresti, A. (2002): *Categorical Data Analysis*. Segunda edición, Wiley, Canadá.
- [2] Burgueño. M. J; García Bastos, J.L y González-Buitrago J.M. (1995): Las curvas ROC en la evaluación de las pruebas diagnósticas. *Medicina Clínica* 104 661-670.
- [3] Cuadras, C. M. (2014): *Nuevos Métodos de Análisis Multivariante*. CMC Editions Manacor Barcelona.
- [4] Hosmer, D. W. Jr., Lemeshow, y S., Rodney X. S. (2013): *Applied Logistic Regression*. Tercera edición, Wiley, Canadá, 2013.
- [5] Johnson, R. y Wichern, D. (2014): *Applied Multivariate Statistical Analysis*. Sexta edición, Pearson Education Limited, N. York.
- [6] Sánchez-Villegas, A; Faulín Fajardo y Fco. Javier. (2006): *Bioestadística Amigable*. Segunda edición, España.
- [7] Sistachs, V. (2005): *Un estudio del modelo de Regresión Logística Binaria bajo el paradigma Bayesiano*. Tesis de Doctorado Universidad de La Habana, Cuba.

INTERACCIONES MEDICAMENTOSAS: UN ESTUDIO DE CORRELACIONES CANÓNICAS EN PACIENTES PEDIÁTRICOS

Gladys Linares Fleites¹, María de Lourdes Sandoval Solís²,
Brenda Catalina Matías Castillo³, Hortensia J. Reyes Cervantes³,
Reyna Almaray⁴ y José A. Ticante Roldán¹

¹ Instituto de Ciencias. ² Facultad de Ciencias de la Computación.

³ Facultad de Ciencias Físico Matemáticas.

⁴ Facultad de Ciencias Químicas.

Benemérita Universidad Autónoma de Puebla
Puebla, Puebla, México.

ABSTRACT

A study on drug interactions in children 2 to 6 years, exploring the relationships between age, weight and height with disease suffering, the types and the number of drugs they had swallowed, was performed. Canonical Correlation Analysis assuming linearity in the relationship between the two sets of quantitative variables observed on the same individuals was used. However, as the correlation between the two sets of variables was low, which would indicate that the relationship is nonlinear, other procedures canonical correlations were used. The kernels methods that transform data using a nonlinear function and send them to a space larger were studied. These procedures gave higher correlations, corroborating that the characteristics of patients and drug interactions are strongly related.

KEYWORDS: canonical correlation with kernels, criterion of independence of Hilbert Schmidt

RESUMEN

Se realizó un estudio sobre interacciones medicamentosas en niños de 2 a 6 años explorando las relaciones existentes entre la edad, el peso y la talla con las enfermedades que padecían, los tipos y el número de medicamentos que habían ingerido. Se utilizó el Análisis de Correlación Canónica suponiendo linealidad en la relación entre los dos conjuntos de variables cuantitativas observadas sobre los mismos individuos. Sin embargo, como la correlación entre los dos conjuntos de variables fue baja, lo que pudiera indicar que la relación no es lineal se utilizaron otros procedimientos de correlaciones canónicas no lineales. Se profundizó en los métodos kernels que transforman los datos mediante una función no lineal y los envían hacia un espacio de mayor dimensión con respecto al espacio en el que se encuentran y una vez que los datos son transformados, se lleva a cabo el Análisis de Correlación Canónico para estudiar la relación entre los dos conjuntos de variables en el nuevo espacio al que fueron transformados. Estos procedimientos dieron correlaciones más altas, corroborando que las características de los pacientes y las interacciones entre medicamentos están fuertemente relacionadas.

1. INTRODUCCIÓN.

Cuando hablamos de interacción medicamentosa nos referimos a la modificación que sufre la acción de un medicamento por la presencia simultánea de otro u otros medicamentos, sustancias fisiológicas y sustancias exógenas no medicamentosas en el organismo, que puede traducirse como la aparición de un *efecto terapéutico o tóxico de intensidad mayor o menor a lo habitual o producto de la interacción previsto* (Linares, et al., 2002).

Las interacciones medicamentosas pueden producir alteraciones en la actividad farmacológica, es decir, que el medicamento sea más o menos eficaz o que produzca efectos adversos no deseados. Por ejemplo, la administración de dos o más medicamentos de manera simultánea o próxima en el tiempo, puede provocar modificaciones en la acción de los mismos. Las interacciones medicamentosas pueden originar cambios cualitativos o cuantitativos en los efectos de los fármacos. La mayoría de estas interacciones son conocidas y por lo tanto previsibles y prevenibles.

Los mecanismos por los cuales se producen estas interacciones son complejos y a menudo hay más de uno implicado. Dada la importancia de estas interacciones, se desarrollan a nivel internacional, actividades de estudio y seguimiento de efectos adversos (Ramos y Olivares, 2010).

En México se ha reportado que más del 70% de las intoxicaciones en la infancia son causadas por medicamentos. Las intoxicaciones en la infancia constituyen una causa común de solicitud de atención médica de urgencia. Su génesis es una mezcla compleja de factores relacionados con el niño y el grupo de edad a que pertenece (el hospedero), con los tóxicos y venenos responsables (el agente) y con las condiciones que determinan la

exposición (el ambiente). El diagnóstico de las intoxicaciones, su tratamiento racional y su prevención se fundamentan en el conocimiento de este complejo: hospedero – agente – ambiente. Las características de este complejo pueden variar en el curso del tiempo o de acuerdo a las condiciones regionales de la población estudiada.

Bajo la hipótesis de que las interacciones entre medicamentos pueden provocar reacciones adversas de complejidad en dependencias de las características de los pacientes, se decidió hacer un estudio sobre las posibles reacciones medicamentosas y las características de pacientes pediátricos en un consultorio de la Ciudad de Puebla, Puebla, México. Se exploraron las relaciones existentes entre la edad, el peso y la talla de niños de 2 a 6 años con las enfermedades que padecían y los tipos y número de medicamentos que habían ingerido.

Para el estudio de las interacciones medicamentosas, el método multivariado de Análisis de Correlaciones Canónicas (ACC) puede ser adecuado dado que éste es generalmente utilizado para investigar la presencia de relaciones en dos conjuntos de variables, determinando la correlación existente entre ellos (Linares, *et al.*, 2004).

La correlación canónica puede abordar una amplia gama de objetivos. Estos objetivos pueden ser cualquiera o todos de los siguientes:

- Valorar si dos conjuntos de variables (con mediciones efectuadas en los mismos sujetos) son independientes uno del otro o, por el contrario, determinar la magnitud de las relaciones que existen entre los dos conjuntos.
- Derivar los pesos o ponderaciones para cada uno de los dos conjuntos de variables, de manera que las combinaciones lineales de cada conjunto estén máximamente correlacionados.
- Explicar la naturaleza de las relaciones que existen entre los dos conjuntos de variables.

El ACC fue desarrollado por Hotelling en 1936 como un procedimiento para evaluar la relación lineal entre dos conjuntos de variables aleatorias. Posteriormente surgieron otros procedimientos que permitían explorar en relaciones no lineales entre los dos conjuntos de variables (Van der Burg, 1988). No es hasta finales de los 90's del siglo pasado que se desarrollaron procedimientos no lineales como generalizaciones de las técnicas clásicas de análisis multivariado.

En este capítulo se exponen las posibilidades que brindan los diferentes procedimientos del ACC para estudiar las interacciones medicamentosas en los pacientes pediátricos del estudio mencionado anteriormente. En el primer apartado se desarrolla el caso clásico y en el segundo las posibilidades gráficas de las correlaciones canónicas no lineales. En el tercer apartado se desarrolla el Análisis de Correlación Canónica Kernel (ACCK), que proyecta los datos hacia un espacio de características de alta dimensión y en ese nuevo espacio se realiza el ACC. Sin embargo, como el ACCK carece de interpretabilidad y robustez frente a características irrelevantes, en el cuarto apartado se introducen las dos extensiones del ACC no lineales que dependen del criterio de independencia de Hilbert-Schmidt (HSIC) y la alineación objetivo del kernel centrado (KTA), que permiten la eliminación de características irrelevantes, mientras extraen combinaciones de características fuertemente asociadas.

2. ANÁLISIS DE CORRELACIÓN CANÓNICA

Cuando se requiere encontrar las correlaciones entre dos conjuntos de variables cuantitativas, el Análisis de Correlación Canónica (ACC) es la técnica estadística que puede usarse. (Johnson, 2007). Mediante combinaciones lineales de las variables observadas, se obtienen nuevas variables ficticias tal que la correlación entre las proyecciones de los datos de estas nuevas variables es máxima.

Se está interesado en las medidas de asociación entre dos grupos de variables. El primer grupo, de p variables, está representado por el vector aleatorio $X^{(1)}$ de orden $(p \times 1)$. El segundo grupo, de q variables, está representado por el vector aleatorio $X^{(2)}$ de orden $(q \times 1)$. Se asume, que $X^{(1)}$ representa el conjunto más pequeño, es decir, $p \leq q$.

Para los vectores $X^{(1)}$ y $X^{(2)}$ conjuntos sea:

$$\begin{aligned} E[X^{(1)}] &= \boldsymbol{\mu}^{(1)}, & Cov(X^{(1)}) &= \boldsymbol{\Sigma}_{11} \\ E[X^{(2)}] &= \boldsymbol{\mu}^{(2)}, & Cov(X^{(2)}) &= \boldsymbol{\Sigma}_{22} \\ Cov(X^{(1)}, X^{(2)}) &= \boldsymbol{\Sigma}_{12} = \boldsymbol{\Sigma}'_{21} \end{aligned}$$

Además se encuentra que el vector

$$X_{[(p+q) \times 1]} = \begin{bmatrix} X^{(1)} \\ \dots \\ X^{(2)} \end{bmatrix}$$

tiene media

$$\boldsymbol{\mu}_{[(p+q) \times 1]} = \begin{bmatrix} \boldsymbol{\mu}^{(1)} \\ \dots \\ \boldsymbol{\mu}^{(2)} \end{bmatrix}$$

y matriz de covarianza

$$\boldsymbol{\Sigma}_{[(p+q) \times (p+q)]} = E(\mathbf{X} - \boldsymbol{\mu})(\mathbf{X} - \boldsymbol{\mu})' = \begin{bmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{bmatrix}$$

Las covarianzas entre cada par de variables de diferente conjunto son contenidas en $\boldsymbol{\Sigma}_{12}$, o equivalentemente en $\boldsymbol{\Sigma}_{21}$. Esto es, los pq elementos de $\boldsymbol{\Sigma}_{12}$ miden la asociación entre los dos conjuntos. La tarea principal del análisis de correlación canónica es resumir las asociaciones entre los conjuntos $X^{(1)}$ y $X^{(2)}$ en términos de las covarianzas (correlaciones) elegidas, más que las pq covarianzas de $\boldsymbol{\Sigma}_{12}$.

Combinaciones lineales proporcionan medidas de resúmenes simples de un conjunto de variables. Sea

$$U = \mathbf{a}'\mathbf{X}^{(1)}$$

$$V = \mathbf{b}'\mathbf{X}^{(2)}$$

donde \mathbf{a} y \mathbf{b} son vectores de coeficientes. Además se tiene que

$$Var(U) = \mathbf{a}'Cov(\mathbf{X}^{(1)})\mathbf{a} = \mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a}$$

$$Var(V) = \mathbf{b}'Cov(\mathbf{X}^{(2)})\mathbf{b} = \mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b}$$

$$Cov(U, V) = \mathbf{a}'Cov(\mathbf{X}^{(1)}, \mathbf{X}^{(2)})\mathbf{b} = \mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}$$

Se requiere buscar los vectores de coeficientes \mathbf{a} y \mathbf{b} tales que:

$$Corr(U, V) = \frac{\mathbf{a}'\boldsymbol{\Sigma}_{12}\mathbf{b}}{\sqrt{\mathbf{a}'\boldsymbol{\Sigma}_{11}\mathbf{a}}\sqrt{\mathbf{b}'\boldsymbol{\Sigma}_{22}\mathbf{b}}} \quad (1)$$

es lo más grande posible.

La *primera pareja de variables canónicas*, es la pareja de combinaciones lineales U_1, V_1 con varianzas unitarias, las cuales maximizan la correlación (1). La *segunda pareja de variables canónicas*, es la pareja de combinaciones lineales U_2, V_2 con varianzas unitarias, las cuales maximizan la correlación (1) entre todas las opciones que no están relacionados con la primera pareja de variables canónicas. Así la *k-ésima pareja de variables canónicas*, es la pareja de combinaciones lineales U_k, V_k con varianzas unitarias, las cuales maximizan la correlación (1) entre todas las opciones que no están relacionados con la $(k-1)$ pareja de variables canónicas.

El problema de las correlaciones canónicas no es más que un problema de extremos condicionados, por lo que puede resolverse aplicando el método de los multiplicadores de Lagrange, aunque la vía más ampliamente utilizada en la actualidad es a través de la descomposición singular de una matriz.

Las correlaciones canónicas pueden ser utilizadas para datos muestrales (Linares, 2006, p. 85). Lo que debe hacerse es sustituir los parámetros poblacionales por los estadísticos muestrales.

Si los datos provienen de una distribución normal multivariada, la matriz de varianzas y covarianzas muestral S es un estimador máximo verosímil de la matriz de varianzas y covarianzas poblacional Σ , luego los valores de las correlaciones canónicas muestrales son estimadores máximo verosímiles de los valores poblacionales correspondientes, excepto cuando hay valores propios de la población repetidos.

2.1 Aplicación del ACC al estudio de reacciones medicamentosas en pacientes pediátricos.

Se realizó un estudio para analizar las posibles causas de las reacciones medicamentosas en pacientes pediátricos de un consultorio de la Ciudad de Puebla, explorando las relaciones existentes entre la edad, el peso y la talla de niños de 2 a 6 años con ciertas enfermedades que han padecido y con los tipos y número de medicamentos que han ingerido (Almaray, 2003).

Se obtuvo una base de datos con información de 292 pacientes pediátricos para estudiar las relaciones entre los grupos de variables presentados en el Cuadro 1.

Cuadro1. Variables del estudio de interacciones medicamentosas

Primer grupo de variables (X)	Segundo grupo de variables (Y)
Edad	Número de enfermedades
Peso	Número de medicamentos
Talla	Número de fármacos

Para determinar el grado de relación entre los dos grupos de variables, se realizó el ACC usando el paquete CCA del software R (González *et al.*, 2008).

Como primera parte, se obtuvieron las correlaciones entre cada variable, éstas correlaciones son presentadas gráficamente en la Figura 1 en donde una correlación alta positiva está representada por el color rojo oscuro, la correlación alta negativa por el color azul fuerte y las correlaciones bajas mediante el color verde. Se observa que la correlaciones entre las variables del primer grupo son altas y positivas (están presentadas en el primer cuadro superior), por otro lado, se observa que la correlación entre las variables del segundo grupo es alta y positiva solo entre la variable número de medicamento y número de fármacos. Para las correlaciones entre las variables del grupo X con el grupo Y las correlaciones son bajas.

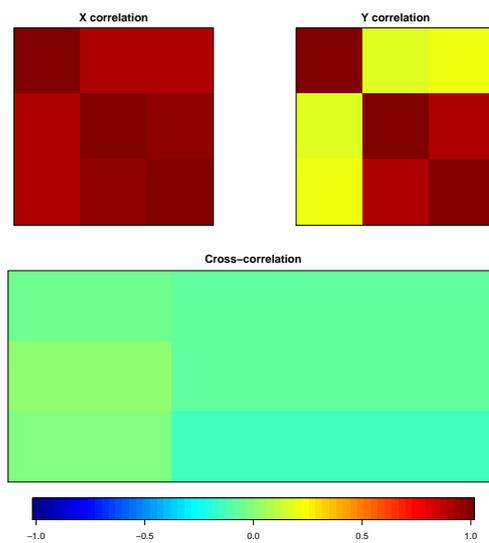


Figura 1. Correlaciones entre las variables.

Los resultados para el ACC clásico están en el Cuadro 2. El valor de la primera correlación canónica obtenida es 0.178, lo que indica que la relación existente entre los dos grupos de variables es baja; de igual forma, la segunda y tercera correlaciones canónicas también son bajas.

Cuadro 2. Correlaciones y vectores canónicos obtenidos para los datos de interacciones medicamentosas

Correlaciones canónicas			
		.178119	.114667
			.020926
Variables canónicas			
Primer conjunto de variables canónicas			
		1	2
ad o la	Ed		3
	Pes	.77529	.47194
	Tal	.00015	0.00030
		0.13047	.01932
			.08139
Segundo conjunto de variables canónicas			
		1	2
ermedades dicamentos macos	Enf		3
	Me	0.241900	1.52956
	Fár	.314396	1.29064
		.521393	.26899
			1.30682

En la Figura 2 se muestran las gráficas de las proyecciones de los datos de las dos primeras variables canónicas. Se observa que los datos se agrupan en líneas horizontales, pero no se observa alguna relación lineal entre las proyecciones.

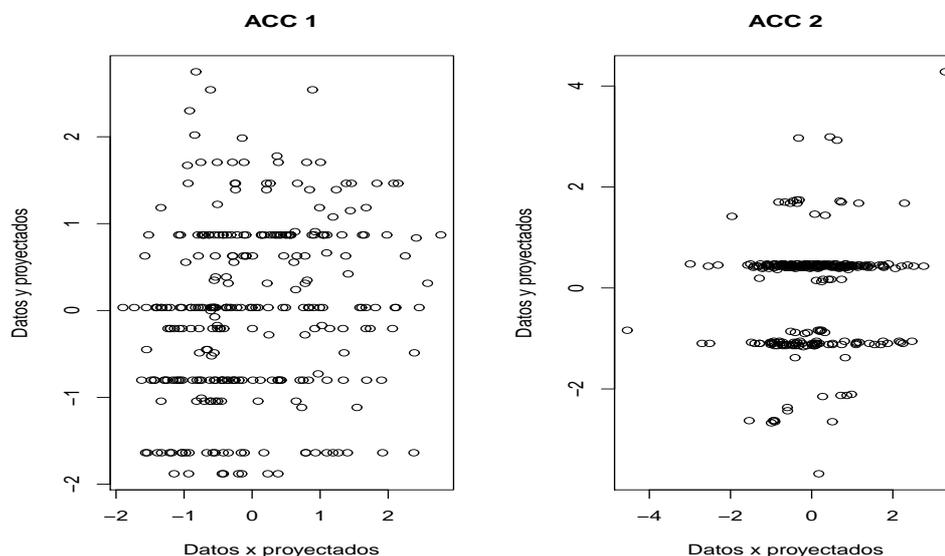


Figura 2. Proyecciones de las variables canónicas

Estos resultados sugieren que la relación entre los dos grupos de variables no es lineal y debía explorarse en este sentido. En Linares, *et al.*, 2004 también se había realizado el ACC con iguales resultados desde el punto de vista exploratorio, pero los resultados inferenciales mostraron que la primera correlación canónica era significativa.

3. Análisis de Correlación Canónica no lineal

El ACC funciona adecuadamente si los datos forman una nube de puntos con forma de hiperelipsoide, pero en la medida que los patrones ideales no se presenten en los datos esta técnica no proporciona resultados adecuados. Para estas situaciones se han propuesto una diversidad de alternativas. Una de estas alternativas es la propuesta por la Universidad de Leiden, Países Bajos y que es utilizada en el programa SPSS versión 21 con el procedimiento OVERALS, 2014.

El propósito del análisis de correlación canónica no lineal es determinar la similitud de dos o más conjuntos de variables (Vander Burg, 1988). Al igual que en el análisis de correlación canónica lineal, el objetivo es explicar el máximo posible de la varianza sobre las relaciones existentes entre dos conjuntos de variables numéricas en un espacio de pocas dimensiones. Inicialmente, las variables de cada conjunto se combinan linealmente de forma que las combinaciones lineales tengan una correlación máxima entre sí. Una vez dadas estas combinaciones, se establece que las combinaciones lineales subsiguientes no estén correlacionadas con las combinaciones anteriores y que también tengan la mayor correlación posible.

A diferencia del análisis de correlación canónica lineal, sin embargo, el análisis de correlación canónica no lineal no asume un nivel de medición de intervalo en las variables, ni asume que las relaciones entre los conjuntos de variables sean lineales.

El análisis de correlación canónica no lineal coincide con el análisis de correlación canónica categórico mediante el escalamiento óptimo. En el enfoque de escalamiento óptimo las variables pueden tener diferentes escalas de medición, esto es, pueden ser nominales, ordinales o numéricas. Por otra parte, se pueden analizar relaciones no lineales entre las variables.

Finalmente, en lugar de maximizar las correlaciones entre los conjuntos de variables, los conjuntos se comparan con un conjunto compromiso desconocido definido por las puntuaciones de los objetos.

3.1 Aplicación OVERALS al estudio de reacciones medicamentosas en pacientes pediátricos.

Como mencionamos anteriormente, el análisis de correlación canónica no lineal se conoce también por el acrónimo OVERALS. En el Cuadro 3 se muestran las saturaciones de las dos primeras dimensiones que brinda este

procedimiento. Obsérvese que en la primera dimensión la más alta saturación (peso o ponderación) se produce en la variable talla del primer grupo de variables, mientras que para el segundo grupo, las variables medicamento y fármaco son las de más alta saturación.

Cuadro 3. Saturaciones en componentes

Conjunto	Dimensión	
	1	2
edad	.425	.117
talla	.627	-.037
peso	.484	-.200
Enfermedad	-.071	-.771
Medicamentos	-.784	-.054
Fármacos	-.743	-.034

En la Figura 3 se muestra una representación gráfica que corrobora la fuerte relación que existe entre estos dos grupos de variables, destacándose la importancia de las variables talla, medicamentos y fármacos, que se encuentran agrupadas en el centro de los ejes.

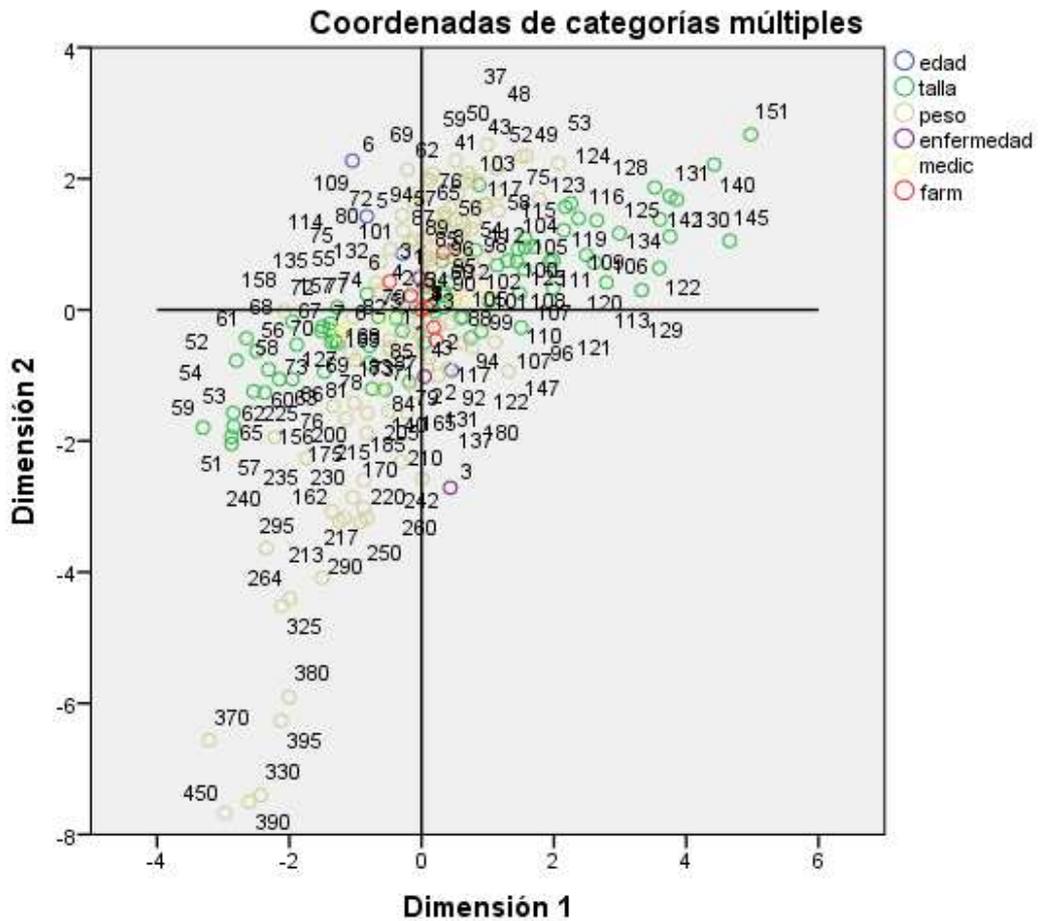


Figura 3. Proyecciones de las dos primeras dimensiones

4. ANÁLISIS DE CORRELACIÓN CANÓNICA CON KERNELS

En la década de los años 90 se introdujo una nueva clase de algoritmos para las técnicas del análisis multivariado, entre las que se encuentran las correlaciones canónicas. Estos algoritmos se basan en la idea, aparentemente paradójica, de transformar a los datos mediante una función no lineal, hacia un espacio de mayor dimensión al espacio en el que se encuentran inicialmente y realizar el análisis multivariado en los datos transformados (Shawe Taylor y Cristianini, 2004). Estos nuevos enfoques dieron origen al Análisis de Correlaciones Canónicas con Kernel (ACCK).

A continuación desarrollamos las ideas esenciales sobre funciones kernel, que constituyen el basamento teórico de estos nuevos algoritmos (Gibaja-Martins, 2010). En esencia un algoritmo basado en funciones kernel para estudiar relaciones entre grupos de variables consiste en:

- Un conjunto de elementos en el que se desea explorar la relaciones entre grupos de variables. Llamaremos a este conjunto *Espacio de Entrada (Input Space)* y lo denotaremos por X .

- Un conjunto de patrones al que llamaremos *Espacio de Características (Feature Space)*, que debe tener una estructura algebraica de *Espacio de Hilbert* y que denotaremos por F . En particular, el espacio F debe estar dotado de un *producto escalar*:

$$\langle \cdot, \cdot \rangle: F \times F \rightarrow \mathbb{R}$$

- Una función $\Phi: X \rightarrow F$ que incrusta (*embeds*) cada elemento del *input space* X en el *feature space* F . Esta función Φ recibe el nombre de *incrustación (embedding)*.

- Un algoritmo de correlaciones canónicas lineales en el *feature space* F que ha sido *kernelizado*, es decir, ha sido rediseñado de manera tal que para alcanzar su objetivo de evaluar la relación en F , el algoritmo no necesita disponer de los valores concretos de $\Phi(x) \forall x \in X$ sino tan solo de los productos escalares entre las imágenes de los elementos de X , es decir, $\langle \Phi(x), \Phi(z) \rangle \forall x, z \in X$, siendo $\langle \cdot, \cdot \rangle: F \times F \rightarrow \mathbb{R}$ el producto escalar definido en F al que se ha aludido previamente.

- Una función real definida sobre el producto cartesiano del *input space* con él mismo $k: X \times X \rightarrow \mathbb{R}$, llamada función *kernel*, que a cada pareja de elementos del *input space* X le hace corresponder el producto escalar en F de sus respectivas imágenes por la función Φ , es decir que:

$$k(x, z) = \langle \Phi(x), \Phi(z) \rangle \quad \forall x, z \in X$$

Este sencillo esquema proporciona una inteligente solución al dilema entre los algoritmos de evaluación de relaciones lineales o no lineales, reuniendo en un nuevo enfoque las mejores características de cada uno.

En este contexto, es necesario hacer algunas definiciones que serán utilizadas en el desarrollo de los algoritmos (Mika, 2002).

Definición. Sean $x_1, x_2, \dots, x_p \in X$ un conjunto de observaciones, la matriz $K_{p \times p}$ con elementos

$$K_{ij} = k(x_i, x_j)$$

es llamada la *matriz de Gram*.

Hasta el momento sólo se tiene una manera de verificar que una función es un kernel, esto es, construyendo un espacio de características donde la función corresponde a la primera representación de la asignación característica y luego calcular el producto interno entre las dos imágenes.

A continuación se introduce un método alternativo de demostrar que una función candidata es un kernel. Esto proporcionará una de las herramientas teóricas necesarias para crear nuevos kernels, y combinar los kernels antiguos para formar otros nuevos. La relación con las matrices semidefinidas positivas es muy importante en este contexto.

Definición. Una matriz simétrica A es **semidefinida positiva** si sus valores propios son todos no negativos. Esto es, se cumple que $v^T A v \geq 0$ para todos los vectores $v \in V$.

De igual manera una matriz simétrica A es **definida positiva** si sus valores propios son positivos o se cumple que $v^T A v > 0$ para todos los vectores $v \in V$ donde $v \neq 0$.

La siguiente proposición se cumple:

Proposición. Las matrices de Gram y kernel son semidefinidas positivas.

Dados los resultados anteriores, se tiene el siguiente teorema de caracterización de funciones kernel.

Teorema: Una función $k: X \times X \rightarrow \mathbb{R}$ que es continua o tiene un dominio finito, puede descomponerse como $k(x, y) = \langle \phi(x), \phi(y) \rangle$ en un mapeo de características ϕ en un espacio de Hilbert F si y sólo si satisface la propiedad de ser finita semidefinida positiva.

Sea k una función kernel semidefinida positiva sobre un conjunto X , se define el espacio de características $H = \{f|f: X \rightarrow \mathbb{R}\}$ y el mapeo $\phi: X \rightarrow \mathbb{R}^X$ como

$$\phi(x) = k(\cdot, x)$$

Se puede probar que el conjunto de todas las combinaciones lineales de la forma

$$f(\cdot) = \sum_{t=1}^p \alpha_t k(\cdot, x_t)$$

para algún $p, \alpha_t \in \mathbb{R}$ arbitrario forman un espacio vectorial. Especialmente para todas las funciones de la forma anterior se tiene que

$$f(x) = \langle k(\cdot, x), f \rangle_F$$

donde $\langle \cdot, \cdot \rangle_F$ denota el producto interno en algún espacio de Hilbert. En particular se tiene

$$\langle k(\cdot, x), k(\cdot, z) \rangle_F = \langle \phi(x), \phi(z) \rangle_H = k(x, z).$$

Definición. Sea X un conjunto no vacío, y F un espacio de Hilbert de funciones $f: X \rightarrow \mathbb{R}$. Entonces F es llamado el espacio de Hilbert reproductor de kernel dotado con el producto interno $\langle \cdot, \cdot \rangle_F$ si existe una función $k: X \times X \rightarrow \mathbb{R}$ con las propiedades siguientes

- k tiene la propiedad de reproducción, $\langle f, k(\cdot, x) \rangle = f(x)$ para toda $f \in F$, en particular, $\langle k(\cdot, x), k(\cdot, z) \rangle = k(x, z)$.
- k se extiende por H , es decir, $H = \overline{\text{span}\{k(\cdot, x) | x \in X\}}$.

Para las diferentes funciones kernel se tiene la siguiente proposición.

Proposición. Sea k_1 y k_2 kernels sobre $X \times X, X \subseteq \mathbb{R}^n, a \in \mathbb{R}^+, f(\cdot)$ una función de valor real en X
 $\phi: X \rightarrow \mathbb{R}^m$

Con k_3 un kernel sobre $\mathbb{R}^m \times \mathbb{R}^m$, y \mathbf{B} una matriz semidefinida positiva simétrica de orden $n \times n$. Entonces las siguientes funciones son kernels

1. $k(x, z) = k_1(x, z) + k_2(x, z)$,
2. $k(x, z) = ak_1(x, z)$,
3. $k(x, z) = k_1(x, z)k_2(x, z)$,
4. $k(x, z) = f(x)f(z)$,
5. $k(x, z) = k_1(\phi(x), \phi(z))$,
6. $k(x, z) = x'Bz$

Existen diferentes tipos de kernels; según Karatzoglou *et al.*, 2004 algunos de ellos son:

- Lineal. $k(x, z) = x \cdot y$.
- Gaussiano. $k(x, z) = \exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$.
- Polinomial. $k(x, z) = (x \cdot y + c)^d$.
- Tangente Hiperbólica. $k(x, z) = \tanh(x \cdot y + c)$.

Teniendo en cuenta las ideas anteriores se ha desarrollado el Análisis de Correlación Canónica con Kernels (ACCK)

Ya vimos que con el Análisis de Correlación Canónica clásico, a veces no se puede evaluar el grado de relación existente entre los grupos de variables debido a su falta de linealidad. El Análisis de Correlación Canónica con Kernel ofrece una solución alternativa por medio de la proyección de los datos en un espacio de características de alta dimensión. A continuación, siguiendo el trabajo de Hardoon *et al.*, 2004 se desarrolla el ACCK.

La matriz de covarianzas puede ser reescrita como las matrices de datos de X y Y , teniendo como vector muestral las filas; estas matrices son de tamaño $m \times n$. Sean

$$C_{xx} = X'X$$

$$C_{xy} = X'Y$$

las matrices de varianzas y covarianzas, reescritas en términos de las matrices de datos de X y Y . Sean W_x y W_y direcciones que pueden reescribirse como la proyección de los datos en las direcciones α y β como sigue

$$W_x = X'\alpha$$

$$W_y = Y'\beta$$

Como en el caso lineal, el objetivo principal del ACCK, es encontrar los vectores canónicos en términos de los coeficientes α y β tal que

$$\rho = \max_{\alpha, \beta} \frac{\alpha' XX' YY' \beta}{\alpha' XX' XX' \alpha \beta' YY' YY' \beta}$$

Sea $K_x = XX'$ y $K_y = YY'$ las matrices kernel (matriz de Gram) correspondientes a los dos representaciones. Entonces sustituyendo de obtiene

$$\rho = \max_{\alpha, \beta} \frac{\alpha' K_x K_y \beta}{\alpha' K_x^2 \alpha \beta' K_y^2 \beta}$$

La ecuación anterior no se ve afectada por el cambio de escala de α y β . Por lo tanto, el problema de optimización de ACCK formulado en la ecuación antes descrita es equivalente a maximizar el numerador sujeto a

$$\begin{aligned} \alpha' K_x^2 \alpha &= 1 \\ \beta' K_y^2 \beta &= 1. \end{aligned}$$

Al igual que en el análisis de correlación canónica clásico, para encontrar los valores de α y β se usa el método de multiplicadores de Lagrange, mismo que conlleva a solucionar un problema de valores y vectores propios.

4.1 Aplicación del ACCK al estudio de reacciones medicamentosas en pacientes pediátricos.

Como se observó anteriormente, las correlaciones canónicas entre los dos conjuntos de variables es baja y podría estar mostrando que la relación no es lineal. Se propone entonces realizar el ACCK a los datos usando los kernel Gaussiano, polinomial y tangente hiperbólica. Éste análisis se llevó a cabo usando el paquete kernlab en R (Karatzoglou *et al.*, 2004). Las correlaciones canónicas son presentadas en el Cuadro 3. Para el kernel gaussiano se usó $\sigma=0.1$, para el kernel tangente hiperbólica se usó $c = 0$; y para el kernel polinomial se utilizaron diferentes valores de c y d , pero no se logró obtener algún resultado

Cuadro 3. Comparación de las correlaciones canónicas obtenidas mediante los diferentes métodos kernel.

Método Kernel	ρ_1	ρ_2	ρ_3
Gaussiano	0.966398	0.931275	0.91657
Tangente Hiperbólica	0.999657	0.530523	0.530523

Como se puede ver en el Cuadro 3, la primera correlación canónica obtenida usando el kernel gaussiano y tangente hiperbólica es alta, 0.966 y 0.999, respectivamente. Esto indica que al transformar los datos se obtuvo una correlación alta.

Observando las proyecciones de las variables canónicas en la Figura 3 de la primera y segunda correlación canónica se observa que los datos representados por el ACCK con kernel gaussiano parecen formar una línea horizontal, lo que indicaría una relación lineal. Los datos representados por el ACCK con kernel tangente hiperbólica para los primeros vectores canónicos se acumulan todos en un círculo en un determinado lugar, mientras que las proyecciones de los datos de los segundos vectores canónicos continúan dentro de un círculo pero están dispersos.

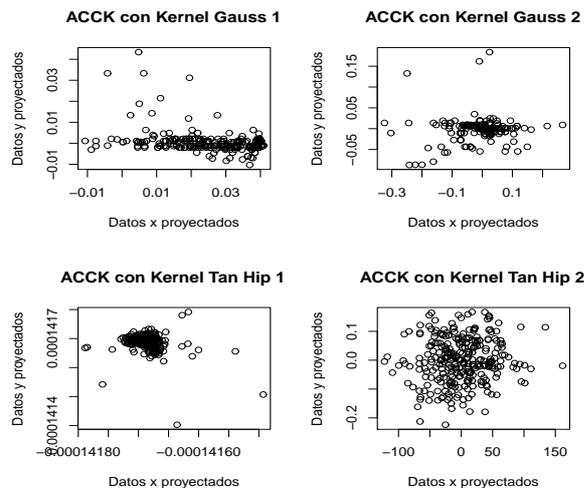


Figura 3. Proyecciones de las variables canónicas usando kernels.

Sin embargo, el procedimiento no permite interpretar los vectores canónicos, sólo brinda las correlaciones canónicas suponiendo no linealidad.

5. ANÁLISIS DE CORRELACIÓN CANÓNICA CON KERNEL USANDO HSIC Y KTA

A continuación se presentan dos métodos que tratan de superar las limitaciones inherentes del ACCK, como son, la falta de interpretabilidad debido a la transformación de los vectores de datos en un espacio de Hilbert abstracto y la incapacidad de eliminar rasgos irrelevantes desde las variables originales (Chang, *et al.*, 2013)

El primer método es el criterio de independencia de Hilbert-Schmidt (HSIC) que es la norma Hilbert-Schmidt cuadrada del operador de covarianza cruzada entre el espacio de probabilidad X y Y . El HSIC se puede utilizar como una medida de independencia cuando se asocia con algún kernel. En general, al igual que el ACCK, en el HSIC se tiene como objetivo calcular los vectores u y v tal que maximicen

$$\rho_h = \frac{1}{(n-1)^2} \text{tr}(K^u \tilde{K}^v)$$

sujeto a $\|u\| = \|v\| = 1$, donde K^u es la matriz de Gram para los datos proyectados $u'x_i$ y \tilde{K}^v la matriz de Gram centrada con ij -ésima entrada.

$$\tilde{K}_{ij}^v = K_{ij}^v - \frac{1}{n} \sum_{i=1}^n K_{ij}^v - \frac{1}{n} \sum_{j=1}^n K_{ij}^v + \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n K_{ij}^v.$$

Por otro lado, se define el método de la alineación objetivo del kernel centrado (KTA), que no es más que la versión normalizada del HSIC, y al igual que éste método, el KTA también es una medida de independencia. El método KTA al igual que el ACCK tiene como objetivo calcular los vectores u y v tal que maximicen

$$\rho_a = \frac{\text{tr}(K^u \tilde{K}^v)}{\sqrt{\text{tr}(K^u \tilde{K}^u) \text{tr}(K^v \tilde{K}^v)}}$$

sujeto a $\|u\| = \|v\| = 1$.

Para resolver los dos métodos anteriores Chang *et al.*, 2013 hacen una optimización mediante el método de gradiente descendente, donde se utiliza el método de Nelder-Mead para la selección del tamaño de paso el cual puede fallar a veces al aumentar el costo (cuando se ve atrapado en mínimos locales).

5.1 Aplicación del HSIC y KTA al estudio de reacciones medicamentosas en pacientes pediátricos.

Como se mencionó anteriormente, el ACCK carece de interpretabilidad debido a la transformación de los vectores hacia un espacio de Hilbert. Usando el paquete hsicCCA en R, de Chang *et al.*, 2013, se realizó el análisis de correlaciones canónicas no lineal usando los métodos HSIC y KTA. Los resultados obtenidos son presentados en el Cuadro 4.

Cuadro 4. Resultados del ACCK usando los métodos HSIC y KTA.

Método	ρ		u			v		
	HsicCCA	ρ_1	0.00276	0.8536	-0.5150	-0.0776	0.2200	0.6615
	ρ_2	0.00165	-0.0815	0.0149	-0.9965	-0.9516	0.3070	0.0088
KtaCCA	ρ_1	0.13038	0.9285	-0.0231	0.3704	-0.5109	-0.6399	-0.5739
	ρ_2	0.10559	-0.3704	0.0063	0.9288	0.8521	-0.2889	-0.4363

Se puede observar que para ρ_1 en los vectores u_1 la variable representativa es la edad en ambos métodos mientras que para el vector v_1 las variables representativas son el número de medicamentos y el número de fármacos para el método HSIC y para el método KTA son las tres variables. Para ρ_2 la variable representativa es la talla para u_2 en ambos métodos y para v_2 la variable representativa es el número de enfermedades.

Las proyecciones son presentadas en la Figura 4. Se observa que los datos proyectados de cada pareja de vectores canónicos se comportan igual, éstos forman semicírculos pero dentro de éstos semicírculos los datos se distribuyen en diferentes líneas.

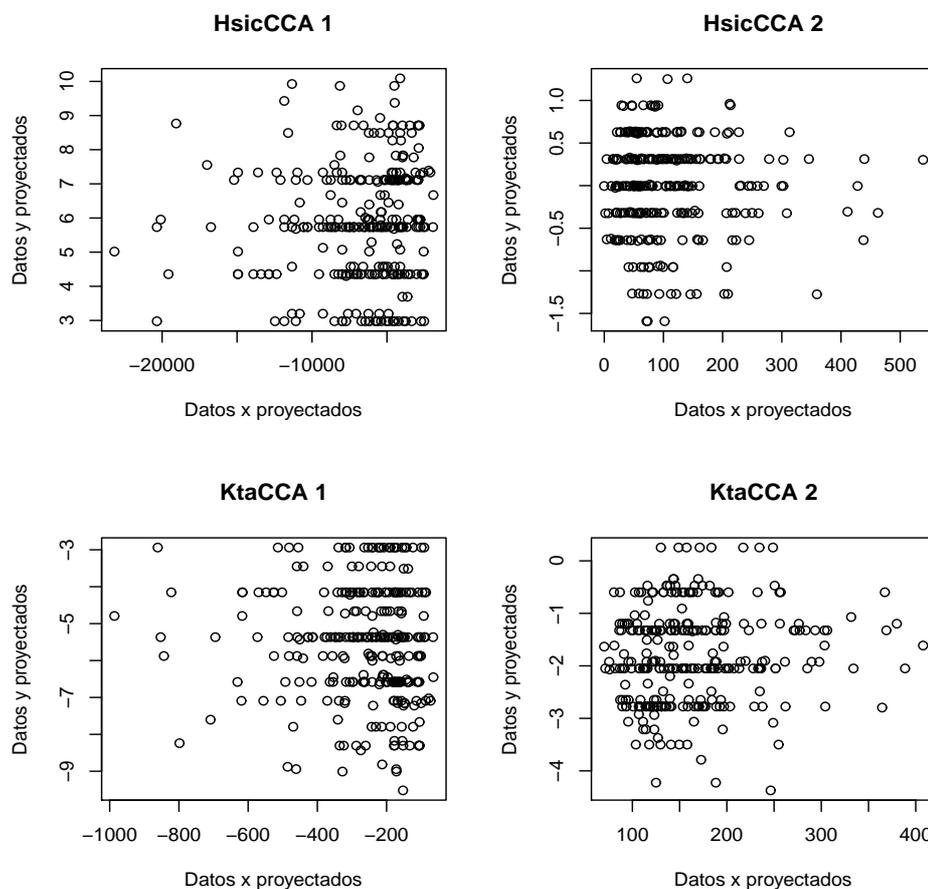


Figura 4. Proyecciones de las variables canónicas usando HSIC y KTA

6. CONCLUSIONES

En la investigación sobre Interacciones Medicamentosas, las diferentes técnicas ideadas utilizando métodos kernels en las correlaciones canónicas, abren nuevas perspectivas para la profundización de estas interacciones y su relación con la calidad de vida de los pacientes.

Utilizando la base de datos generada en Almaray, 2003, para analizar las posibles causas de las reacciones medicamentosas en pacientes pediátricos 2 a 6 años de un consultorio privado de la Ciudad de Puebla, México, se logró evaluar las relaciones existentes entre la edad, el peso y la talla de niños con ciertas enfermedades y con los tipos y número de medicamentos ingeridos por ellos. Los métodos ACC, OVERALS, ACCK, KTA y HSIC se utilizaron y se analizaron los resultados mostrados en cada uno de ellos.

El ACCK con la introducción del KTA y HSIC logró dar una valoración más objetiva de la fuerte relación existente entre las características de los pacientes estudiados y las reacciones, muchas veces adversas, hacia ciertos medicamentos y fármacos interactuando de manera compleja con ciertas enfermedades.

Como puede apreciarse, por su importancia en las aplicaciones (en particular, la referida a las Interacciones Medicamentosas) las correlaciones canónicas como técnica multivariada ha generado y sigue generando mucho trabajo de investigación.

REFERENCIAS

- [1] Almaray, R del C. (2003). Estudio de interacciones medicamentosas en pacientes pediátricos de 2-6 años en un consultorio privado. Tesis de Master en Farmacia Clínica. Facultad de Ciencias Químicas. Benemérita Universidad Autónoma de Puebla, México.
- [2] Chang, B., Kruger, U., Kustra, R. y Zhang, J. (2013). "Canonical Correlation Analysis based on Hilbert-Schmidt Independence Criterion and Centered Kernel Target Alignment". 30th International Conference on Machine Learning. Atlanta, Georgia.
- [3] Gibaja - Martíns JJ (2010). Aprendizaje Estadístico con Funciones Kernel. Tesis de Maestría. UNED. España.
- [4] González, I., Déjean, S., Martin, P. y Baccini, A. (2008). CCA: An R Package to Extend Canonical Correlation Analysis. *Journal in Statistical Software*, 23 (12).
- [5] Haroon, D., Sandor, S. and Shawe-Taylor, J. (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods. *Neural Computation* 16, 2639 - 2664.
- [6] Hotelling, H. (1936). "Relations between Two Sets of Variables". *Biometrika*, 28(3), p. 321-377.
- [7] Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis*. 6th edition. Prentice Hall. New Jersey.
- [8] Karatzoglou, A., Smola, A. y Hornik, K. (2004). kernlab: An S4 Package for Kernel Methods in R. *Journal in Statistical Software*, 11(9).
- [9] Linares, G., Ticante, J.A. y Almaray, R. (2004). Estudio de interacciones medicamentosas en pacientes pediátricos con técnicas multivariadas. Memorias en extenso del 2do. Congreso Nacional de Ciencias de la Computación. Facultad de Computación, BUAP.
- [10] Linares, A, Milián, P., Jiménez, L., Chala, J., Alemán, J., Betancourt, B.Y., Rodríguez, J.M. y Martín, L.M. (2002). Interacciones Medicamentosas. *Acta Farm. Bonaerense* 21 (2): 139-48.
- [11] Linares, G. (2006). *Análisis de Datos Multivariados*. México: Editorial de la Benemérita Universidad de Puebla.
- [12] Mika, S. (2002). Kernel Fisher Discriminants. Tesis de Doctorado. Universidad Técnica de Berlín.
- [13] OVERALS (2014). Data Theory Scaling System Group (DTSS). Faculty of Social and Behavioral Sciences Leiden University. The Netherlands.
- [14] Ramos G., Olivares G. (2010). *Uso racional de medicamentos: una tarea de todos*. Ministerio de Salud. Santiago de Chile.
- [15] Shawe Taylor, J. y Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. First Edition. Cambridge University Press.
- [16] Van der Burg, E. (1988). *Nonlinear canonical correlation and some related techniques*. Leiden: DSWO Press.

ESTIMACIÓN DE RIESGO CARDIOVASCULAR SEGÚN MODELO DE FRAMINGHAM Y MODELO LOGIT EN PACIENTES DE LA JURISDICCIÓN SANITARIA 5 DE GUERRERO.

Agustín Santiago Moreno*, José Maclovio Sautto Vallejo*, Diego Guzmán-Balderas*, Arturo de León Chapa* & Jhony Alberto De La Cruz Vargas**.

*Universidad Autónoma de Guerrero- UAGro,

**Escuela de Medicina y de la Maestría de Salud Pública de la Universidad Peruana Unión.

ABSTRACT

Cardiovascular diseases are the main cause of death in adults around the world. Mexico, as well as other Latin American countries, suffer this phenomena and the prevalence of the factors that originate them are very high, therefore the aim of this research is to estimate the cardiovascular risk, according to two models the Framingham and the logistic regression in patients from the sanitary jurisdiction number five, located in Zihuatanejo, Guerrero, to identify the segment of the population with high cardiovascular risk (HCR). A cross-sectional, non-experimental and explanatory study was made, with a sample of 172 patients older than 20 years, in which factors associated to cardiovascular risk were measured and those two cardiovascular risk diagnostic models were evaluated. The variables considered in the first model (Framingham) were hypertension = T, elevated cholesterol or low density lipoproteins = C, high density lipoproteins = H, diabetes mellitus type II= D, gender = (Male, Female) and age = A. The original model considers the condition of the smoker, although in our data this variable was not included. Due to the origin of the Framingham model, the population studied had differences with the North American population, for that reason the binary logistic regression model was evaluated with the purpose of identifying the variables statistically significant for the classification of patients in cardiovascular risk, considering additional variables such as obesity= OB and physical inactivity = PI. According to the Framingham score, participants qualify as HCR patients, with a score of 0.2 or higher, which corresponds to 10.5% of the population, while the logistic regression model classifies as CR to 15.7% of the population with a correct classification scale of 91.3%. Considering this difference in the classification, the logistic regression model is better for the population studied, due to the smaller number of variables it uses for the diagnosis.

KEYWORDS: cardiovascular risk, Framingham score, logistic regression

RESUMEN

Las enfermedades cardiovasculares son la principal causa de muerte en adultos en el mundo. México, al igual que otros países latinoamericanos, padece este fenómeno y la prevalencia de los factores que las originan son muy altas, por lo que el objetivo de esta investigación fue estimar el riesgo cardiovascular, según dos modelos el de Framingham y el de regresión logística en pacientes de la jurisdicción sanitaria número 5, con cabecera en Zihuatanejo, Guerrero, para identificar el segmento de la población en alto riesgo cardiovascular (ARC). Se realizó un estudio transversal, no experimental y explicativo con una muestra de 172 pacientes mayores de 20 años, en los cuales se midieron factores asociados al riesgo cardiovascular y se evaluaron esos dos modelos de diagnóstico de riesgo cardiovascular. Las variables consideradas en el primer modelo (Framingham) fueron hipertensión = T, colesterol elevado o lipoproteína de baja densidad = C, lipoproteína de alta densidad = H, diabetes mellitus tipo II = D, sexo = (Hombre, Mujer) y edad = E. El modelo original considera la condición de fumador del paciente, aunque en nuestros datos esta variable no fue incluida. Debido al origen del modelo de Framingham, la población estudiada tiene diferencias con la población norteamericana, por lo que se evaluó un modelo de regresión logística binaria con la finalidad de identificar las variables estadísticamente significativas para la clasificación de pacientes en riesgo cardiovascular, considerando variables adicionales tales como obesidad = OB e inactividad física = IF. Según la puntuación de Framingham clasifican en ARC los sujetos cuyas puntuaciones sean 0.2 o mayor, que corresponde al 10.5% de la población, mientras que el modelo de regresión logística clasifica como RC al 15.7% de la población con una tasa de clasificaciones correctas del 91.3%. Aún con esta diferencia en la clasificación, el modelo de regresión logística es mejor para la población estudiada porque utiliza un número menor de variables para el diagnóstico.

1. INTRODUCCIÓN.

Las enfermedades cardiovasculares son la principal causa de muerte en adultos en el mundo (Molinero, 2003), (Ruiz, Segura, & Agustí, 2012), (OMS, 2015). El riesgo cardiovascular se define como la probabilidad de muerte cardiovascular de una persona en un periodo determinado (10 años) (Gobierno Federal, SALUD, SEDENA Y SEMAR), (Alcocer, Lozada, Fanghanel, Sánchez-Reyes, & Campos-Franco,

2011). Existen diversos tipos de enfermedades cardiovasculares, entre ellas podemos mencionar la hipertensión arterial, enfermedad arterial coronaria, enfermedad valvular cardíaca, accidente cerebrovascular (trombosis o derrame cerebral) y fiebre reumática o enfermedad cardíaca reumática (DMEDICINA.com, Salud y bienestar, 2015). Cada año, aproximadamente 37 millones de personas en el mundo sufren un evento cardiovascular (ataque al corazón o una enfermedad vascular cerebral) (Alejandro, 2008O), (López, 2013) y alrededor de 17.5 millones de personas mueren por dichas causas (OMS, 2015). Con frecuencia las enfermedades cardiovasculares resultan de la interacción de diferentes factores de Riesgo Cardiovascular, en donde se incluyen hipertensión, niveles elevados de lípidos en la sangre, diabetes, obesidad y tabaquismo (Ruiz, Segura, & Agusti, 2012). Por tanto, los principales factores de riesgo cardiovascular son: Presión arterial alta o hipertensión arterial (HTA), colesterol elevado o lipoproteína de baja densidad (LDL), las lipoproteínas de alta densidad (HDL, del inglés High density lipoprotein), diabetes, obesidad y sobrepeso, tabaquismo, inactividad física, género, herencia y la edad (Alcocer, Lozada, Fanghänel, Sánchez-Reyes, & Campos-Franco, 2011).

Según afirmación de la OMS, la mayoría de las enfermedades cardiovasculares se pueden prevenir cuando el tratamiento se enfoca de manera global a todos los factores de riesgo que presenta una persona, por ejemplo, tratar la hipertensión arterial al mismo tiempo que la elevación del colesterol; el tener alta la presión arterial o el colesterol por separado representa un riesgo para la persona, pero si están juntos el riesgo se incrementa significativamente. Por lo que la Organización Mundial del Corazón (WHO, 2013), señala de acuerdo a los estudios realizados, que el tratamiento de los factores de riesgo se sigue enfocando en el manejo por separado de cada uno de ellos, en lugar de ser enfocado a la atención del riesgo global del individuo (WHO, 2003).

1.1.1 Prevalencia de RC en México.

En México las enfermedades cardiovasculares son la principal causa de muerte y la prevalencia de los factores que las originan son muy altas. Según datos de la Secretaría del salud de nuestro país y de la encuesta nacional de salud y nutrición 2012, el 30% de la población mayor de 20 años en México es hipertensa (QUO, 2012), Campos-Nonato, Hernández-Barrera, Rojas-Martínez, Pedroza-Tobías, Medina-García, & Barquera, (2013) encontraron que el 31.5% de los adultos mexicanos de 20 años a más padecen de hipertensión, donde el 47.3% desconocían tener esta enfermedad

Alcocer, Lozada, Fanghänel, Sánchez-Reyes, & Campos-Franco (2011), realizaron un estudio con una población de 5803 pacientes, tomando una muestra de 1990 sujetos, trabajadores del hospital General de México supuestamente sanos, excluyendo a pacientes con diabetes e hipertensos y encontraron que el 93.12% clasificó en bajo riesgo (1853 pacientes), el 6.68% en riesgo medio (133 pacientes) y un .020% en alto riesgo (4 pacientes).

La literatura médica divide los factores de riesgo cardiovascular en dos categorías: principales y contribuyentes. Los principales factores de riesgo son aquellos cuyo efecto de aumentar el riesgo cardiovascular ha sido comprobado. Los factores contribuyentes son aquellos que los médicos piensan que pueden dar lugar a un mayor riesgo cardiovascular pero cuyo papel exacto no ha sido definido aún (Farmacias ABC, 2006).

Los principales factores de riesgo cardiovascular son: Presión arterial alta o hipertensión arterial (T), colesterol elevado o lipoproteína de baja densidad (C), las lipoproteínas de alta densidad (H), diabetes (D), obesidad (OB) y sobrepeso (Sp), tabaquismo (Tab), inactividad física (IF), sexo (Hombre, Mujer), herencia (Her) y la edad (E) (Alcocer, Lozada, Fanghänel, Sánchez-Reyes & Campos-Franco, 2011).

Por lo tanto, entre más factores de riesgo tenga una persona, mayores serán sus probabilidades de padecer una enfermedad del corazón. Algunos factores de riesgo pueden cambiarse, tratarse o modificarse y otros no. Pero el control del mayor número posible de factores de riesgo, mediante cambios en el estilo de vida y/o medicamentos, puede reducir el riesgo cardiovascular (Farmacias ABC, 2006).

Los factores contribuyentes de riesgo cardiovascular son: el estrés, hormonas sexuales, anticonceptivos orales y el alcohol, aunque para este último los estudios demuestran que el riesgo cardiovascular es menor en las personas que beben cantidades **moderadas** de alcohol que en las personas que no beben. Según los expertos, el consumo moderado es un promedio de una o dos bebidas por día para los hombres y de una bebida por día para las mujeres (Texas Heart Institute, 2015)

Según los datos de la ENSANUT 2012, estamos ante un verdadero problema de salud que día a día va creciendo, y que tanto las autoridades de salud como los usuarios de los servicios, debemos tomar conciencia de la realidad y hacer una verdadera medicina preventiva para enfrentar el problema (ENSANUT,

2012). Hasta el momento de la consulta, los resultados no están disponibles a nivel de entidad federativa o municipio, por lo que no tenemos un referente actual para los resultados obtenidos en el estudio realizado con pacientes del municipio de Zihuatanejo, Guerrero.

En el presente trabajo se restringe a la descripción del segmento de la población que se encuentra en riesgo vascular, clasificado en esa situación por los “factores de riesgo” manifiestos en ellos, que ahora se sabe desempeñan un papel importante en las probabilidades de que una persona padezca de una enfermedad del corazón y desde la perspectiva médica, hacer las recomendaciones pertinentes para el control de la enfermedad y disminución del riesgo.

2. OBJETIVOS.

2.1 Objetivo general

Comparar la evaluación del riesgo cardiovascular, según el modelo Framingham y el de regresión logística en pacientes de la jurisdicción sanitaria número 5, con cabecera en Zihuatanejo, Guerrero

2.2 Objetivos específicos.

- Estimar la probabilidad de riesgo cardiovascular en la población de pacientes de la jurisdicción sanitaria número 5, según el modelo Framingham.
- Estimar la probabilidad de riesgo cardiovascular en la población de pacientes de la jurisdicción sanitaria número 5, según el modelo de Regresión Logística
- Comparar los métodos anteriores en la estimación de riesgo cardiovascular.

3. MATERIALES Y MÉTODOS.

3.1 Características del estudio.

Se realizó un estudio transversal, no experimental y explicativo con objeto de analizar variables asociadas a la obesidad y diabetes, dentro del programa “Cruzada contra, sobrepeso, Obesidad y Diabetes” del departamento de investigación, en coordinación con el departamento de enfermedades crónicas de los Servicios Estatales de Salud en Guerrero. Se analiza una muestra de 172 pacientes de la jurisdicción sanitaria número cinco con sede en Zihuatanejo, Guerrero, considerada como muestra piloto. Esta muestra se constituyó con personas mayores de 20 años que acudieron a algún tipo de servicio médico y se realizaron pruebas a partir de muestras de sangre sobre perfil de lípidos, glucosa capilar, medidas antropométricas como peso, talla, circunferencia de la cintura e índice aterogénico.

Para el perfil de lípidos se utilizó un aparato portátil cardio Chek, cuyos resultados se obtuvieron en cinco minutos, la glucosa capilar se midió con un glucómetro accu-Chek Performa, que da resultados en diez segundos con alta efectividad. Las medidas antropométricas se hicieron con una báscula con escalímetro, para talla y peso, una calculadora médica para medir riesgo aterogénico, de acuerdo a los resultados de los lípidos y con el monitor de grasa e índice de masa corporal, mediante el método de impedancia bioeléctrica, se estimaba el riesgo y grado de obesidad del paciente, cuyos datos se habían almacenado previamente.

A partir de estas variables y la clasificación de la OMS sobre obesidad, diabetes y riesgo cardiovascular, se obtienen las estimaciones de prevalencia, medidas de asociación, estimación de probabilidades para clasificación y cuyo error estándar se estimó mediante la técnica de remuestreo denominada bootstrap, toda vez que no se trata de una muestra aleatoria, en el sentido que corresponda a algún diseño de muestreo aleatorio establecido en la teoría de muestreo, debido a que, aunque existe un listado de usuarios del servicio médico, la selección aleatoria no se hizo de ese marco de muestreo, sino que se incluyeron en la muestra de estudio a pacientes mayores de 20 años que solicitaron algún servicio médico, en orden de llegada, durante el mes de septiembre de 2012.

3.2 Modelo de Framingham.

Aunque existen gran número de trabajos relativos al estudio de los riesgos de enfermedad cardiovascular, el conocido como estudio de Framingham constituye un pilar básico, y en diferentes formas es ampliamente utilizado para la toma de decisiones terapéuticas en base a la estimación de riesgo proporcionada por el modelo al introducir las características de riesgo del paciente concreto (Rodríguez, 2014).

Los países europeos se caracterizan por una baja incidencia de eventos cardiovasculares respecto al lugar de origen del método (Framingham, Massachusetts, USA), en razón de ello, el modelo de Framingham produce una sobreestimación del riesgo absoluto de enfermedad cardiovascular, lo que podía influir al utilizar ese modelo en la decisión de tratar un exceso de pacientes en países como España o Italia, en base a una sobreestimación del riesgo real. Como es lógico esta inquietud crea la necesidad de desarrollar un modelo más adecuado para este entorno, y así recientemente se ha publicado un trabajo correspondiente a la estimación del riesgo de desarrollar en 10 años una enfermedad cardiovascular fatal en países de Europa (proyecto SCORE) (Molinero, 2003).

Existen otros modelos además de los dos citados, entre los que cabe quizás destacar otro trabajo anterior que incluye pacientes de Norte América y Europa, basado en el proyecto INDANA (Individual Data Analysis of Antihypertensive Intervention Trials), y que también dispone de una calculadora on-line de riesgo (Gueyffier, y otros, 1997) citado por (Molinero, 2003).

Es claro que estos estudios se han realizado en individuos que viven en países desarrollados, en los cuales el estilo de vida urbano industrial propicia el desarrollo de distintos padecimientos asociados a un aumento de la incidencia y prevalencia de las enfermedades cardiovasculares. Para el caso de América latina, en el 2010 se condujo un estudio sobre los factores de riesgo cardiovascular en siete ciudades (Barquisimeto en Venezuela, Bogotá en Colombia, Buenos Aires en la Argentina, Lima en Perú, México DF en México, Quito en Ecuador y Santiago de Chile en Chile) denominado CARMELA (Cardiovascular Risk Factor Multiple Evaluation in Latin America).

Los hallazgos de este estudio no fueron más allá de un estudio descriptivo de la población estudiada y la clasificación de la población en riesgo cardiovascular se realizó utilizando los puntajes de riesgo de Framingham. En el presente trabajo se utilizó este método de clasificación de riesgo cardiovascular, bajo el supuesto de hábitos alimenticios similares, debido a que aunque no se ha demostrado que sea apropiado para la población latinoamericana, no se ha validado aún otro método de puntaje para la estimación de la probabilidad de riesgo cardiovascular en pacientes Mexicanos.

A continuación se describe brevemente el método y las fórmulas a partir de las cuales se obtienen las puntuaciones.

3.2.1 Descripción del modelo Framingham.

En la propuesta original del modelo, Molinero (2003), mencionó que las variables que intervienen son: sexo (H, M), Edad en años, el colesterol sérico en mg/dl (C), fracción de colesterol ligado a lipoproteínas de alta densidad HDL (H), presión arterial (T) sistólica y diastólica, diabetes (D) (sí, no) y fumador (F) (sí, no). Medidas estas variables se calculan las siguientes expresiones:

$$\text{Para hombres } L_H = b_{E1} * \text{Edad} + b_C + b_H + b_T + b_D + b_F$$

$$\text{Para mujeres } L_M = b_{E1} * \text{Edad} + b_{E2} * \text{Edad}^2 + b_C + b_H + b_T + b_D + b_F$$

Los coeficientes b_i para son hombres y mujeres diferentes y se obtienen a partir de la tabla 1, realizando los cálculos correspondientes. Una vez calculado L, se le resta la cantidad G, que es la función evaluada para los valores medios de las variables en estudio que son diferentes para hombres o mujeres, como se muestra a continuación

$$G_{\text{Hombres}} = 3.0975$$

$$G_{\text{Mujeres}} = 9.92545$$

Aplicando la exponencial a la diferencia de $L - G$ se obtiene la función B.

$$B = \exp(L - G)$$

Determinando el valor de la expresión $1 - S^B$, donde S es la función de supervivencia base a 10 años, siendo diferente para hombres y mujeres

$$S_{\text{Hombres}} = 0.90015$$

$$S_{\text{Mujeres}} = 0.96246$$

Por tanto el modelo Framingham para hombre y mujeres es

$$Y_{\text{Hombres}} = 1 - 0.90015^B$$

$$Y_{\text{Mujeres}} = 1 - 0.96246^B$$

donde Y es la probabilidad de padecer un evento cardiovascular en los próximos 10 años.

Tabla 1: Coeficientes para el modelo Framingham (Colesterol total)

Coeficiente	Hombres	Mujeres	Coeficiente	Hombres	Mujeres
$b_{E1} \times \text{edad}$	0.04826	0.33766	b_T Tensión arterial mmHg		
$b_{E2} \times \text{edad}^2$	0	-0.00268	$P_{AS} < 120, PAD < 80$	-0.00226	-0.53363
b_c colesterol			$P_{AS} < 130, PAD < 85$	0	0
<160	-0.65945	-0.26138	$P_{AS} < 140, PAD < 90$	0.2832	-0.06773
160-199	0	0	$P_{AS} < 160, PAD < 100$	0.52168	0.26288
200-239	0.17692	0.20771	$P_{AS} > 160, PAD \geq 100$	0.61859	0.46573
240-279	0.50539	0.24385	b_D Diabetes		
≥ 280	0.65713	0.53513	NO	0	0
b_H HDL mg/dl			SI	0.42839	0.59626
<35	0.49744	0.84312	b_F Fuma		
35-44	0.2431	0.37796	NO	0	0
45-49	0	0.19785	SI	0.52337	0.29246
50-59	-0.05107	0			
≥ 60	-0.4866	-42951			

Fuente: The Framingham Heart Study

3.3 Modelo de regresión logística.

En nuestro caso, ya que nos interesa identificar el segmento de la población que se encuentra en situación crítica de riesgo cardiovascular, realizamos la clasificaron en dos grupo, el primer grupo toma el valor 1, es decir, presencia de la condición y el segundo grupo toma el valor 0, o ausencia de la condición. Las covariables o variables explicativas asociadas al riesgo cardiovascular consideradas en el estudio fueron: hipertensión = T, colesterol elevado o lipoproteína de baja densidad = C, lipoproteína de alta densidad = H, diabetes mellitus tipo II = D, obesidad = OB, inactividad física = IF, sexo = (Hombre, Mujer) y edad = E.

Por lo anterior, tenemos una variable respuesta dicotómica y variables explicativas también dicotómicas, lo que posibilita la formulación de un modelo estadístico para estimar una variable respuesta discreta (binaria) en función de una o varias variables explicativas que podrían ser cuantitativas o cualitativas.

En el caso más simple se quiere explicar una variable aleatoria de respuesta binaria Y, con dos posibles categorías de respuesta (Y_1, Y_2) en función de una variable no aleatoria cuantitativa o cualitativa X.

Si representamos a las dos categorías de Y por los valores 0 y 1, Y tiene distribución de Bernoulli de esperanza

$$E[Y] = P[Y = 1] = p \quad (0 < p < 1)$$

Entonces, la distribución de Y en cada valor observado de X es también Bernoulli de esperanza

$$E[Y|X = x] = P[Y = 1|X = x] = p(x) \quad (1)$$

y varianza

$$Var[Y|X = x] = E[Y^2|X = x] - (E[Y|X = x])^2 = p(x)[1 - p(x)]$$

De este modo, $p(x)$ representa la dependencia de la probabilidad de respuesta 1 respecto de los valores de la variable explicativa.

Si denotamos por $Y(x)$ a la distribución de Y condicionada a $X = x$ ($Y|X = x$), el paso siguiente es construir un modelo adecuado para $Y(x)$ de la forma

$$Y(x) = \text{función}(\text{parámetros}, x, \text{error})$$

Teniendo en cuenta lo razonado anteriormente, buscamos un modelo de la forma

$$Y(x) = F(\alpha + \beta x) + e(x)$$

con $e(x)$ variable aleatoria independientes de esperanza cero, o equivalentemente

$$p(x) = F(\alpha + \beta x) \quad (2)$$

donde F es una función de distribución estrictamente creciente, que a su vez puede expresarse en la forma (Hosmer & Lemeshow, 1999).

$$F^{-1}(p(x)) = \alpha + \beta x$$

El modelo de regresión logística simple es de la forma

$$p(x) = \frac{e^{\alpha + \beta x}}{1 + e^{\alpha + \beta x}} = \frac{1}{1 + e^{-(\alpha + \beta x)}} \quad (3)$$

El modelo se puede escribir equivalentemente en la forma

$$\ln \left[\frac{p(x)}{1 - p(x)} \right] = \alpha + \beta x$$

donde la transformación $\ln[p(x)/(1 - p(x))]$ recibe el nombre de logit y $p(x)/(1 - p(x))$ representa la ventaja (odds) de respuesta 1 para el valor observado x .

3.3.1 Modelo logit múltiple

Consideremos ahora el caso de R variables explicativas cuantitativas no aleatorias (X_1, X_2, \dots, X_R) . Para cada combinación de valores observados $X_1 = x_1, X_2 = x_2, \dots, X_R = x_R$ de las variables explicativas, la variable respuesta Y tiene distribución de Bernoulli

$$Y|(X_1 = x_1, X_2 = x_2, \dots, X_R = x_R) \sim B(1, p(x_1, x_2, \dots, x_R))$$

siendo

$$p(x_1, x_2, \dots, x_R) = P[Y = 1|X_1 = x_1, \dots, X_R = x_R] = E[Y|X_1 = x_1, \dots, X_R = x_R].$$

El modelo de regresión logística múltiple para la variable respuesta binaria Y en términos de valores observados $X_1 = x_1, \dots, X_R = x_R$ de las variables explicativas es de la forma

$$Y(x_1, x_2, \dots, x_R) = p(x_1, x_2, \dots, x_R) + e(x_1, x_2, \dots, x_R),$$

donde $e(x_1, x_2, \dots, x_R)$ son errores aleatorios que se consideran centrados e independientes, de modo que

$$p(x_1, x_2, \dots, x_R) = \frac{e^{(\alpha + \sum_{r=1}^R \beta_r x_r)}}{1 + e^{(\alpha + \sum_{r=1}^R \beta_r x_r)}}.$$

Denotando a partir de ahora $\alpha = \beta_0$, $X = (X_0, X_1, \dots, X_R)'$ y $x = (x_0, x_1, \dots, x_R)'$ con $X_0 = 1$, el modelo quedará resumido como sigue:

$$p(x) = \frac{e^{\sum_{r=0}^R \beta_r x_r}}{1 + e^{\sum_{r=0}^R \beta_r x_r}} = \frac{e^{\beta'x}}{1 + e^{\beta'x}} \quad (4)$$

donde β es el vector columna de parámetros $(\beta_0, \beta_1, \dots, \beta_R)'$.

Equivalentemente el modelo de regresión logística múltiple se puede ver como un modelo de regresión lineal múltiple para la transformación logit

$$\ln \left[\frac{p(x)}{1 - p(x)} \right] = \sum_{r=0}^R \beta_r x_r.$$

3.3.1.1 Interpretación de los parámetros.

- Si $\beta_r = 0 \forall r = 1, \dots, R$ entonces $p(x) = e^{\beta_0}/(1 + e^{\beta_0})$ que quiere decir que la variable Y es independiente de las variables explicativas.
- β_0 es el valor común del logaritmo de las ventajas de respuesta $Y = 1$ frente a respuesta $Y = 0$ cuando la respuesta es independiente de las variables explicativas.

Por otro lado, β_0 es el valor del logaritmo de la ventaja de respuesta $Y = 1$ para un individuo con $X_1 = X_2 = \dots = X_R = 0$.

- El cociente de ventajas de respuesta $Y = 1$ para dos combinaciones diferentes de valores de las variables explicativas, $x_1 = (1, x_{11}, \dots, x_{1R})'$ y $x_2 = (1, x_{21}, \dots, x_{2R})'$, es de la forma

$$\theta(x_1, x_2) = \frac{\frac{p(x_1)}{1-p(x_1)}}{\frac{p(x_2)}{1-p(x_2)}} = \frac{e^{\sum_{r=0}^R \beta_r x_{1r}}}{e^{\sum_{r=0}^R \beta_r x_{2r}}} = e^{\sum_{r=1}^R \beta_r (x_{1r} - x_{2r})}.$$

Para dos valores x_1 y x_2 que se diferencien en una unidad, $x_{1r} - x_{2r} = 1 \forall r = 1, \dots, R$ se tendría

$$\theta(\Delta X_1 = 1, \dots, \Delta X_R = 1) = e^{\sum_{r=1}^R \beta_r} = \prod_{r=1}^R e^{\beta_r}.$$

- Para dar una interpretación más intuitiva de los parámetros del modelo, vamos a calcular el cociente de ventajas de respuesta 1 cuando una de las variables explicativas se incrementa en una unidad y las otras se controlan haciendo que tomen valores fijos.

3.3.2 Ajuste de los modelos logit.

En esta sección vamos a abordar el problema de la estimación de los parámetros de los modelos logit.

Los datos están constituidos por una muestra de tamaño N de la variable aleatoria respuesta Y . Es decir, se dispone de N observaciones de N variables de Bernoulli independientes (sucesión de unos y ceros), a cada una de las cuales corresponde una determinada combinación de niveles, (x_0, x_1, \dots, x_R) de las R variables explicativas X_1, X_2, \dots, X_R .

Si denotamos por $x_q = (x_{q0}, x_{q1}, \dots, x_{qR})$ ($q = 1, \dots, Q$) a la q -ésima combinación de valores de las R variables explicativas en la muestra, pueden ocurrir dos casos:

- Para cada individuo muestral existe una combinación diferente de niveles de las R variables explicativas ($Q = N$). Esto significa que hay una única observación de la variable aleatoria de respuesta Y en cada combinación de valores de las variables explicativas, y suele ocurrir cuando las variables explicativas son todas continuas.
- A individuos muestrales diferentes corresponden valores iguales de las variables explicativas ($Q < N$). Esto quiere decir que hay más de una observación de la variable aleatoria de respuesta en cada combinación de valores de las variables explicativas.

Denotando por n_q al número de observaciones muestrales con $X = x_q$ y por y_q al número de respuestas $Y = 1$ de entre estas n_q observaciones, se dispone de una muestra de Q variables aleatorias independientes Y_q con distribuciones $B(n_q, p_q)$, donde $p_q = P[Y = 1 | X = x_q]$. Por lo tanto, $E[Y_q] = n_q p_q$ y $\sum_{q=1}^Q n_q = N$.

Observemos que Y_q representa en general el número de respuestas $Y = 1$ en cada $X = x_q$. En el caso en que no hay observaciones repetidas ($Q = N$) los valores observados y_q corresponden a las respuestas binarias individuales (1 o 0).

El modelo de regresión logística muestral es de la forma

$$p_q = \frac{e^{\sum_{r=0}^R \beta_r x_{qr}}}{1 + e^{\sum_{r=0}^R \beta_r x_{qr}}} \quad (5)$$

En forma lineal

$$L_q = \ln \left[\frac{p_q}{1 - p_q} \right] = \sum_{r=0}^R \beta_r x_{qr},$$

3.3.2.1 Estimación por máxima verosimilitud.

Se sabe que los estimadores de máxima verosimilitud (MV) son los valores de los parámetros que dan máxima probabilidad (verosimilitud) a los datos observados. Para encontrarlos tendremos que maximizar la función de verosimilitud de los datos respecto de los parámetros del modelo logit.

La función de verosimilitud es el producto de las Q funciones de distribución de probabilidad de las Q binomiales independientes Y_q dada por

$$\prod_{q=1}^Q \binom{n_q}{y_q} p_q^{y_q} (1 - p_q)^{n_q - y_q},$$

cuyo núcleo (que alcanza el máximo en el mismo punto) es

$$\prod_{q=1}^Q p_q^{y_q} (1 - p_q)^{n_q - y_q} = \left(\prod_{q=1}^Q (1 - p_q)^{n_q} \right) \left(\prod_{q=1}^Q \left(\frac{p_q}{1 - p_q} \right)^{y_q} \right).$$

Maximizaremos como es usual el logaritmo del núcleo de la función de verosimilitud que bajo el modelo de regresión logística (5) es de la forma

$$\begin{aligned} L(\beta) &= \sum_{q=1}^Q n_q \ln(1 - p_q) + \sum_{q=1}^Q y_q \ln \left(\frac{p_q}{1 - p_q} \right) \\ &= - \sum_{q=1}^Q n_q \ln \left(1 + e^{\sum_{r=0}^R \beta_r x_{qr}} \right) + \sum_{q=1}^Q y_q \left(\sum_{r=0}^R \beta_r x_{qr} \right) \\ &= - \sum_{q=1}^Q n_q \ln \left(1 + e^{\sum_{r=0}^R \beta_r x_{qr}} \right) + \sum_{r=0}^R \left(\sum_{q=1}^Q y_q x_{qr} \right) \beta_r, \end{aligned}$$

que depende de los datos observados y_q sólo a través de los estadísticos suficientes $\sum_{q=1}^Q y_q x_{qr}$.

Derivando respecto a cada uno de los parámetros β_r se obtiene

$$\frac{\partial L(\beta)}{\partial \beta_r} = \sum_{q=1}^Q y_q x_{qr} - \sum_{q=1}^Q n_q x_{qr} \left(\frac{e^{\sum_{r=0}^R \beta_r x_{qr}}}{1 + e^{\sum_{r=0}^R \beta_r x_{qr}}} \right).$$

Igualando a cero se obtienen las ecuaciones de verosimilitud

$$\sum_{q=1}^Q y_q x_{qr} - \sum_{q=1}^Q n_q \hat{p}_q x_{qr} = 0 \quad r = 0, 1, \dots, R,$$

donde \hat{p}_q es el estimador MV de p_q

$$\hat{p}_q = \frac{e^{\sum_{r=0}^R \beta_r x_{qr}}}{1 + e^{\sum_{r=0}^R \beta_r x_{qr}}} \quad (6)$$

y los $\hat{\beta}_r$ los estimadores MV de los parámetros.

Como la log-verosimilitud es una función cóncava para los modelos logit, los estimadores MV de sus parámetros existen y son únicos excepto en ciertos casos. Supongamos que ordenamos las observaciones según una única variable explicativa, y que a todos los valores para los que la variable respuesta es cero corresponden valores de la variable explicativa menores que aquellos para los que la respuesta es uno. Este caso se conoce como separación completa y lleva a la no existencia de los estimadores MV. Está demostrado que para que existan los estimadores MV tiene que darse cierto traslape en los datos. Cuando las variables explicativas son continuas es difícil encontrar separación completa en las observaciones. Un estudio más profundo sobre la separación completa y la existencia de los estimadores MV se puede ver en Ryan (1997).

Observemos finalmente que las ecuaciones de verosimilitud no son lineales en los parámetros por lo que requieren el uso de métodos de solución iterativa como el de Newton-Raphson que no se estudiará en este trabajo, pero que se puede consultar en la literatura.

3.3.2.2 Propiedades de los estimadores MV

Wald en 1943 proporcionó resultados asintóticos generales para los estimadores de máxima verosimilitud demostrando que, bajo ciertas condiciones de regularidad, los estimadores de máxima verosimilitud tienen distribución asintótica normal con media el valor poblacional del parámetro estimado y matriz de covarianzas dada por la inversa de la matriz de información de Fisher. Esto significa que los estimadores de máxima verosimilitud son asintóticamente insesgados y además se puede hacer inferencia sobre ellos basándose en dicha distribución normal cuando el tamaño muestral es suficientemente grande.

En nuestro caso, el estimador MV del vector de parámetros β del modelo de regresión logística múltiple converge en distribución a la siguiente distribución normal

$$\hat{\beta} \sim N(\beta, Cov(\hat{\beta})) \text{ cuando } N \rightarrow \infty$$

Donde la matriz de varianzas y covarianzas es

$$Cov(\hat{\beta}) = \left(-E \left[\frac{\partial^2 L}{\partial \beta_r \partial \beta_s} \right]\right)^{-1} = (X' \text{Diag}[n_q p_q (1 - p_q) X])^{-1} \quad (7)$$

Por lo tanto, las raíces cuadradas de los elementos de la diagonal de esta matriz son los errores estándar de los estimadores de los parámetros del modelo.

La estimación MV de esta matriz de covarianzas se obtiene sustituyendo en la expresión anterior p_q por su estimación MV \hat{p}_q .

3.3.3 Inferencia en regresión logística.

Una vez estimados los parámetros del modelo de regresión logística múltiple nos proponemos hacer inferencia para extrapolar los resultados muestrales a la población.

3.3.3.1 Contrastes de bondad de Ajuste.

Recordemos que y_q representa el número respuestas $Y = 1$ (éxitos) en las n_q observaciones correspondientes a la q -ésima combinación de valores de las variables explicativas.

Una vez estimados los parámetros, a partir de ellos se estiman los logit, y a partir de estos las probabilidades \hat{p}_q . Por lo tanto, las frecuencias esperadas de respuesta $Y = 1$, estimadas bajo el modelo de regresión logística, son en este caso de la forma $\hat{m}_q = n_q \hat{p}_q$.

Para contrastar la bondad del ajuste global del modelo cuando el número de observaciones n_q en cada combinación de valores de las variables explicativas es grande se dispone del estadístico chi-cuadrado de Pearson y del estadístico de Wilks de razón de verosimilitudes. Cuando n_q no es suficientemente grande se usará el estadístico de Hosmer y Lemeshow que es una versión modificada del estadístico chi-cuadrado de Pearson.

El test global de bondad de ajuste del modelo de regresión logística múltiple contrasta la hipótesis nula

$$H_0: p_q = \frac{e^{\sum_{r=0}^R \beta_r x_{qr}}}{1 + e^{\sum_{r=0}^R \beta_r x_{qr}}} \quad \forall q = 1, \dots, Q$$

frente a la hipótesis alternativa

$$H_1: p_q \neq \frac{e^{\sum_{r=0}^R \beta_r x_{qr}}}{1 + e^{\sum_{r=0}^R \beta_r x_{qr}}} \text{ para algún } q.$$

Prueba chi-cuadrado de Pearson.

El estadístico chi-cuadrado de Pearson de bondad de ajuste a un modelo M de regresión logística de la forma (6) es

$$\chi^2(M) = \sum_{q=1}^Q \frac{(y_q - n_q \hat{p}_q)^2}{n_q \hat{p}_q (1 - \hat{p}_q)} = \sum_{q=1}^Q \frac{n_q (y_q - \hat{m}_q)^2}{\hat{m}_q (n_q - \hat{m}_q)}$$

donde \hat{p}_q es la estimación MV de p_q bajo el modelo M y $\hat{m}_q = n_q \hat{p}_q$ es la estimación de MV de los valores esperados $m_q = n_q p_q$.

Este estadístico se obtiene como la suma de los estadísticos chi-cuadrado de bondad de ajuste a cada una de las distribuciones $B(n_q, p_q)$ que generan a los datos muestrales bajo la hipótesis nula de que las probabilidades p_q verifiquen el modelo M . Es decir,

$$\chi^2(M) = \sum_{q=1}^Q \left[\frac{(y_q - n_q \hat{p}_q)^2}{n_q \hat{p}_q} + \frac{((n_q - y_q) - (n_q - n_q \hat{p}_q))^2}{(n_q - n_q \hat{p}_q)} \right]$$

$$= \sum_{q=1}^Q \frac{n_q (y_q - n_q \hat{p}_q)^2}{n_q \hat{p}_q (n_q - n_q \hat{p}_q)} = \sum_{q=1}^Q \frac{(y_q - n_q \hat{p}_q)^2}{n_q \hat{p}_q (1 - \hat{p}_q)}$$

Este estadístico tiene distribución asintótica chi-cuadrado con grados de libertad obtenidos como la diferencia entre el número de parámetros p_q (transformaciones logit muestrales) y el número de parámetros independientes en el modelo. Es decir,

$$\chi^2(M) \sim \chi_{Q-(R+1)}^2 \quad \text{cuando } n_q \rightarrow \infty$$

Por lo tanto, se rechazará la hipótesis nula H_0 al nivel de significación α cuando se verifica que

$$\chi^2(M)_{obs} \geq \chi_{Q-(R+1)}^2; \alpha,$$

siendo $\chi_{Q-(R+1)}^2$ el cuantil de orden $(1 - \alpha)$ que es el valor de la distribución $\chi_{Q-(R+1)}^2$ que acumula a su derecha probabilidad α . Equivalentemente, si se define el p-valor del contraste como la probabilidad acumulada a la derecha del valor observado

$$p_{valor} = P[\chi^2(M) \geq \chi^2(M)_{obs}],$$

se rechazará la hipótesis nula con nivel de significación α cuando se verifique que $p_{valor} \leq \alpha$.

Prueba chi-cuadrado de razón de verosimilitudes.

El estadístico de Wilks de razón de verosimilitudes para el contraste de bondad de ajuste de un modelo de regresión logística múltiple M es de la forma

$$G^2(M) = 2 \left[\sum_{q=1}^Q (n_q - y_q) \ln \left(\frac{n_q - y_q}{n_q - \hat{m}_q} \right) + \sum_{q=1}^Q y_q \ln \left(\frac{y_q}{\hat{m}_q} \right) \right].$$

El estadístico de Wilks de razón de verosimilitudes se obtiene como menos dos veces el logaritmo del cociente entre el supremo de la verosimilitud bajo la hipótesis nula y el supremo de la verosimilitud en la población. Es decir, si denotamos por V a la función de verosimilitud de los datos

$$G^2(M) = -2 \ln \left(\frac{\text{Sup}_{\{p \text{ bajo } H_0\}} V(p)}{\text{Sup}_{\{p\}} V(p)} \right) = 2 \left[\text{Sup}_{\{p\}} L(p) - \text{Sup}_{\{p \text{ bajo } H_0\}} L(p) \right],$$

donde L es la log-verosimilitud. Se sabe que

$$\text{Sup}_{\{p \text{ bajo } H_0\}} L(p) = L(\hat{p})$$

donde es el vector cuyas componentes son las estimaciones MV, \hat{p}_q , de las probabilidades p_q bajo el modelo M . Por otro lado, derivando la log-verosimilitud respecto de cada p_q e igualando a cero, se demuestra fácilmente que

$$\text{Sup}_{\{p\}} L(p) = L(f)$$

donde f es el vector $Q \times 1$ cuyas componentes son las proporciones observadas de respuesta $Y = 1$ en cada combinación de valores de las variables explicativas, definidas por $f_q = y_q/n_q$. Observemos que $L(f)$ corresponde al máximo de la log-verosimilitud bajo el *modelo saturado* que es aquel que se ajusta perfectamente a los datos (las frecuencias esperadas de respuesta $Y = 1$ bajo este modelo coinciden con las observadas, $\hat{m}_q = n_q f_q = y_q$ y tiene tantos parámetros libres como observaciones diferentes de las variables explicativas (transformaciones logit diferentes en nuestro caso). Un ejemplo simple de modelo saturado es el de regresión lineal simple con sólo dos datos.

Para simplificar denotaremos al máximo de la log-verosimilitud bajo un modelo por L_M . En particular, L_S representará el máximo de la log-verosimilitud bajo el modelo saturado.

Recordemos que la log-verosimilitud de los datos es

$$L(p) = \sum_{q=1}^Q \ln \binom{n_q}{y_q} + \sum_{q=1}^Q y_q \ln(p_q) + \sum_{q=1}^Q (n_q - y_q) \ln(1 - p_q).$$

Por lo tanto, operando adecuadamente se tiene la expresión del estadístico $G^2(M)$ en términos de las frecuencias esperadas bajo el modelo $\hat{m}_q = n_q \hat{p}_q$

$$G^2(M) = 2(L_S - L_M) = 2(L(f) - L(\hat{p}))$$

$$= 2 \left[\sum_{q=1}^Q (n_q - y_q) \ln \left(\frac{n_q - y_q}{n_q - \hat{m}_q} \right) + \sum_{q=1}^Q y_q \ln \left(\frac{y_q}{\hat{m}_q} \right) \right]$$

El estadístico de Wilks de razón de verosimilitudes tiene distribución asintótica chi-cuadrado con grados de libertad obtenidos como la diferencia entre la dimensión del espacio paramétrico y la dimensión de este espacio bajo la hipótesis nula. En el caso del modelo de regresión logística múltiple el número de grados de libertad es la diferencia entre el número de parámetros p_q y el número de parámetros β_r bajo el modelo. En total $Q - (R + 1)$ grados de libertad

$$G^2(M) \sim \chi_{Q-(R+1)}^2 \text{ cuando } n_q \rightarrow \infty$$

Por lo tanto, se rechazará la hipótesis nula H_0 al nivel de significación α cuando se verifica que

$$G^2(M)_{\text{obs}} \geq \chi_{Q-(R+1)}^2; \alpha,$$

o bien cuando el p-valor del contraste sea menor o igual que el nivel de significación

$$p\text{-valor} = P[G^2(M) \geq G^2(M)_{\text{obs}}] < \alpha.$$

El estadístico $G^2(M)$ suele ser llamado *devianza* y juega un papel similar al de la suma de los cuadrados de los residuos en regresión lineal.

Observemos que los estadísticos de bondad de ajuste de Pearson y de razón de verosimilitudes tienen la misma distribución asintótica chi-cuadrado con los mismos grados de libertad. Para que $\chi^2(M)$ y $G^2(M)$ tengan distribuciones chi-cuadrado aproximadas debe ocurrir que el número de observaciones n_q en cada nivel observado de las covariables (variables explicativas) sea grande. Por ello, en el caso de covariables continuas estos contrastes de bondad de ajuste no tienen distribución aproximada chi-cuadrado debido a que lo usual es obtener observaciones distintas de las variables explicativas para cada individuo de la muestra. La teoría asintótica se aplica de forma más natural a modelos logísticos con variables explicativas categóricas. En el caso de variables continuas los valores observados se agrupan en intervalos que definan una partición del rango de valores de las variables explicativas dando lugar al test de Hosmer y Lemeshow que se presenta a continuación.

Prueba de Hosmer y Lemeshow

Cuando no hay un número suficiente de observaciones n_q en cada combinación de valores x_q de las variables explicativas, no se puede asumir la distribución chi-cuadrado de los estadísticos de Pearson y de razón de verosimilitudes como buena. La norma para poder usar estos contrastes es que el 80% de las frecuencias estimadas bajo el modelo, $\hat{m}_q = n_q \hat{p}_q$, sean mayores que cinco y todas mayores que uno.

El estadístico de Hosmer y Lemeshow es el estadístico chi-cuadrado de Pearson de bondad de ajuste al modelo después de agrupar adecuadamente los datos en intervalos, de modo que su valor depende fuertemente del número de clases resultantes de la agrupación.

Supongamos que agrupamos las variables explicativas obteniendo como resultado G grupos o clases, denotando por n'_g al número total de observaciones en el g-ésimo grupo, por u_g al número de respuestas $Y = 1$ en el g-ésimo grupo, y por \bar{p}_g a la probabilidad estimada bajo el modelo de respuesta $Y = 1$ para el g-ésimo grupo obtenida como la media de las probabilidades \hat{p}_q de los valores x_q de dicho grupo. Entonces, el estadístico de Hosmer y Lemeshow es de la forma

$$C = \sum_{g=1}^G \frac{(u_g - n'_g \bar{p}_g)^2}{n_g \bar{p}_g (1 - \bar{p}_g)},$$

y tiene también distribución asintótica chi-cuadrado con $G - 2$ grados de libertad.

El problema en la aplicación práctica de este estadístico es la selección de las clases. Se sabe que si el número de clases G es menor que seis, este contraste lleva casi siempre a aceptar el modelo como adecuado (potencia muy baja). Sin embargo usar muchos grupos puede llevar a tamaños muestrales pequeños en cada grupo.

Hosmer y Lemeshow (1985) aconsejan $G = 10$ construyendo las clases en base a los deciles de las probabilidades estimadas bajo el modelo.

Medidas globales de bondad de ajuste.

Para cuantificar la bondad del ajuste global del modelo se dispone de medidas como la *Tasa de clasificaciones correctas* (TCR) y medidas *R-cuadrado* alternativas al coeficiente R^2 de la regresión lineal.

a) Tasa de clasificaciones correctas.

La tasa de clasificaciones correctas es la proporción de individuos clasificados correctamente por el modelo, obtenida como el cociente entre el número de observaciones clasificadas correctamente y el tamaño muestral N .

Un individuo es clasificado correctamente por el modelo logit cuanto su valor observado de respuesta (1 o 0) coincide con su valor estimado por el modelo. Para asignar respuesta $Y = 1$ o $Y = 0$ bajo el modelo a los datos se elige un punto de corte (cut-off), $p \in (0,1)$, de modo que a una observación con valor $X = x_q$ se le estima respuesta $Y = 1$ si $\hat{p}_q \geq p$ y se le estima respuesta $Y = 0$ cuando $\hat{p}_q < p$.

Como punto de corte para clasificar en unos y ceros a las observaciones se suele elegir 0.5 aunque es más apropiado elegir la proporción de unos en la muestra. El programa SPSS permite elegir el punto de corte de manera que calcula la tasa de clasificaciones correctas para una partición muy fina de puntos de corte en $(0,1)$ de modo que el analista de datos pueda elegir como más apropiado el punto de corte que proporciona la máxima tasa de clasificaciones correctas.

b) Medidas tipo R^2

Es conocido que en regresión R^2 da la reducción proporcional en varianza entre la varianza condicional de la respuesta y la varianza marginal, describiendo la magnitud de la asociación entre el predictor lineal y la respuesta. Para el caso de datos de respuesta categórica se han definido medidas análogas a R^2 pero ninguna de ellas es tan útil como este coeficiente. Siguiendo esta filosofía el coeficiente R^2 de bondad de ajuste global del modelo de regresión logística se podría definir como la siguiente medida de reducción proporcional en el error:

$$1 - \frac{\sum_{q=1}^Q (y_q - n_q \hat{p}_q)^2}{\sum_{q=1}^Q (y_q - \bar{y})^2}$$

que en el caso de ajustar el modelo de probabilidad lineal por mínimos cuadrados, coincide con el cuadrado del coeficiente de correlación lineal entre $\{y_q\}$ y $\{\hat{m}_q\}$.

Una desventaja importante de esta última medida es que no tiene en cuenta en la estructura del error la dependencia de la varianza de Y_q respecto de p_q . Además, dado un modelo y un conjunto de datos, los valores estimados de los parámetros que maximizan esta medida no son los estimadores MV ni tampoco estimadores eficientes. Además, al calcularla con estimadores MV, esta medida puede decrecer al añadir una variable explicativa al modelo. También se ha demostrado que este coeficiente puede llegar a tomar valores pequeños cuando el ajuste es casi perfecto. Estos problemas se podrían corregir definiendo una medida que pondere las desviaciones cuadráticas por el inverso de las varianzas estimadas pero la medida resultante es sin duda más complicada de cara a la interpretación.

Teniendo en cuenta que en el caso del modelo de regresión lineal el coeficiente de determinación se podía definir de forma equivalente a partir del cociente entre el máximo de la verosimilitud bajo el modelo nulo dado sólo por un término constante (V_0) y el máximo de la verosimilitud bajo el modelo ajustado con todos los parámetros (V_M), en regresión logística se calcula R^2 de la siguiente forma (ver Ryan (1997)):

$$R_{LN}^2 = 1 - \left(\frac{V_0}{V_M} \right)^{\frac{2}{N}},$$

que recibe el nombre de R^2 de Cox y Snell y es implementado por SPSS.

Al aumentar el número de parámetros de un modelo, aumenta el máximo de la verosimilitud, y por otro lado, como las probabilidades están entre cero y uno, la medida R^2 toma valores entre 0 y 1.

Aunque esté acotada entre cero y uno, la medida R^2 así calculada no toma necesariamente el valor uno como máximo sino

$$\text{máx} R_{LN}^2 = 1 - (V_0)^{\frac{2}{N}}.$$

Además este valor puede ser próximo a cero cuando hay pocos datos por lo que se propone como medida de bondad de ajuste el siguiente coeficiente de determinación ajustado:

$$R_N^2 = \frac{R_{CN}^2}{\text{máx}R_{CN}^2},$$

que recibe el nombre de R^2 de Nagelkerke y es implementado también por el programa SPSS.

3.3.4 Contraste de los parámetros del modelo.

En este apartado pretendemos contrastar si un subconjunto de los parámetros β_r del modelo de regresión logística, que denotaremos por $\gamma = (\gamma_1, \dots, \gamma_l)'$, es nulo. Por lo tanto la hipótesis nula del contraste es

$$H_0: \gamma = 0.$$

3.3.4.1 Contrastes de Wald

Están basados en la normalidad asintótica de los estimadores de máxima verosimilitud. Consideremos el contraste de hipótesis

$$\begin{aligned} H_0: \gamma &= 0 \\ H_1: \gamma &\neq 0. \end{aligned}$$

El estimador MV de γ , denotado por $\hat{\gamma}$, tiene distribución normal asintótica de media γ y matriz de covarianzas estimada $\widehat{Cov}(\hat{\gamma})$ obtenida de forma inmediata a partir de la matriz de covarianzas estimada para todos los parámetros $\widehat{Cov}(\hat{\beta})$ correspondiente a la expresión (7).

Como consecuencia el estadístico de Wald de este contraste es la siguiente forma cuadrática

$$\hat{\gamma}' [\widehat{Cov}(\hat{\gamma})]^{-1} \hat{\gamma},$$

que bajo la hipótesis nula tiene distribución chi-cuadrado asintótica con l grados de libertad (número de parámetros nulos bajo la hipótesis nula).

Por lo tanto, se rechazará la hipótesis nula al nivel de significación α cuando el valor observado de este estadístico sea mayor o igual que el cuantil de orden $(1 - \alpha)$ de la distribución χ_l^2 , denotado por $\chi_l^2; \alpha$.

Si en un determinado modelo M se quiere contrastar la igualdad a cero de un solo parámetro

$$\begin{aligned} H_0: \beta_r &= 0 \\ H_1: \beta_r &\neq 0, \end{aligned}$$

el estadístico del contraste será

$$W = \frac{\hat{\beta}_r^2}{\hat{\sigma}^2(\hat{\beta}_r)},$$

que bajo la hipótesis nula tiene distribución chi-cuadrado asintótica con un grado de libertad, por ser el cuadrado de una normal estándar.

Por lo tanto se rechazará la hipótesis nula al nivel de significación α cuando se verifique que

$$W_{Obs} \geq \chi_1^2; \alpha.$$

3.3.4.2 Contrastes condicionales de razón de verosimilitudes.

Supongamos que un modelo de regresión logística M_G se ajusta bien y queremos contrastar si un subconjunto de sus parámetros $\gamma = (\gamma_1, \dots, \gamma_l)$ son nulos. Denotemos por M_p al modelo más simple que resulta al hacer cero estos parámetros en M_G , de modo que el modelo particular M_p está anidado en el modelo general M_G . Las hipótesis de este contraste se pueden expresar como

$$\begin{aligned} H_0: \gamma &= 0 (M_p \text{ se verifica}) \\ H_1: \gamma &\neq 0 (\text{asumiendo cierto } M_G) \end{aligned}$$

Asumiendo que M_G se verifica, el estadístico de la prueba de razón de verosimilitudes para contrastar si M_p se verifica es de la forma:

$$\begin{aligned} G^2(M_p|M_G) &= -2(L_p - L_G) \\ &= -2(L_p - L_S) - (-2(L_G - L_S)) = G^2(M_p) - G^2(M_G), \end{aligned}$$

donde L_S , L_P y L_G son los máximos de la log-verosimilitud bajo la suposición de que se verifican los modelos saturado, M_P y M_G , respectivamente. Esto quiere decir que la prueba de razón de verosimilitudes para contrastar dos modelos anidados es la diferencia de los contrastes de razón de verosimilitudes de bondad de ajuste para cada modelo.

Además, el estadístico $G^2(M_P|M_G)$ tiene, bajo el modelo M_P (hipótesis nula), distribución chi-cuadrado con grados de libertad igual a la diferencia entre los grados de libertad de las distribuciones chi-cuadrado asintóticas de $G^2(M_P)$ y $G^2(M_G)$, que coincide con el número de parámetros que se anulan bajo H_0 . En este caso,

$$G^2(M_P|M_G) \rightarrow \chi_l^2.$$

Este estadístico tiene el mismo papel en regresión logística que el numerador de la prueba F parcial en regresión lineal.

Igual que con el estadístico de Wald, se rechazará la hipótesis nula al nivel de significación α cuando el valor observado $G_{obs}^2(M_P|M_G)$ sea mayor o igual que el cuantil de orden $(1 - \alpha)$ de la distribución χ_l^2 , denotado por $\chi_l^2; \alpha$.

Para el contraste sobre un parámetro de hipótesis nula $H_0: \beta_r = 0$, se construye con la misma filosofía el estadístico $G^2(M_P|M_G)$ que en este caso tiene distribución asintótica χ_1^2 .

Existen casos en los que la prueba de Wald no es tan potente como la prueba de razón de verosimilitudes, proporcionando a veces resultados no deseables. Por ello se aconseja usar la prueba de razón de verosimilitudes en los procedimientos de selección de variables.

3.3.5 Intervalos de confianza.

En esta sección se construirán intervalos de confianza aproximados basados en la distribución normal asintótica de los estimadores MV.

3.3.5.1 Intervalos de confianza para los parámetros.

Vamos a construir un intervalo de confianza (I.C.) aproximado con nivel de confianza $1 - \alpha$ para cada uno de los parámetros β_r del modelo de regresión logística múltiple. Para ello recordemos que la distribución asintótica de $\hat{\beta}_r$ es $N(\beta_r, \hat{\sigma}^2(\hat{\beta}_r))$.

Entonces se tiene que

$$P \left[-z_{\alpha/2} \leq \frac{\hat{\beta}_r - \beta_r}{\hat{\sigma}(\hat{\beta}_r)} \leq z_{\alpha/2} \right] = 1 - \alpha,$$

de donde se obtiene el siguiente intervalo de confianza aproximado para β_r al nivel $(1 - \alpha)$:

$$\hat{\beta} \pm z_{\alpha/2} \hat{\sigma}(\hat{\beta}_r).$$

3.3.5.2 Intervalos de confianza para las transformaciones logit.

Sea la transformación logit

$$L(x) = \ln \left[\frac{p(x)}{1 - p(x)} \right] = x' \beta,$$

donde $p(x) = P[Y = 1|X = x]$ con $x = (1, x_1, \dots, x_R)'$ y β el vector de parámetros del modelo de regresión logística múltiple.

Para construir un I.C. aproximado para $L(x)$ nos basaremos en la distribución normal asintótica de su estimador MV definido por

$$\hat{L}(x) = x' \hat{\beta}.$$

A partir de la distribución normal de $\hat{\beta}$ se obtiene que $\hat{L}(x)$ tiene distribución asintótica normal de media $L(x)$ y varianza estimada $\hat{\sigma}^2(\hat{L}(x)) = x' \widehat{Cov}(\hat{\beta}) x$.

Como consecuencia, un I.C. aproximado para el logit poblacional $L(x)$ al nivel $1 - \alpha$ es de la forma

$$\hat{L}(x) \pm z_{\alpha/2} \hat{\sigma}(\hat{L}(x)).$$

3.3.5.3 Intervalo de confianza para las probabilidades de respuesta $Y = 1$.

Haciendo uso de la transformación: $\hat{p}(x) = \frac{e^{\hat{L}(x)}}{1+e^{\hat{L}(x)}}$ se obtiene el siguiente I.C. aproximado al nivel de confianza $1 - \alpha$ para la probabilidad $p(x)$ de respuesta $Y = 1$ dado el valor $X = x$:

$$\frac{e^{(\hat{L}(x) \pm z_{\alpha/2} \hat{\sigma}(\hat{L}(x)))}}{1 + e^{(\hat{L}(x) \pm z_{\alpha/2} \hat{\sigma}(\hat{L}(x)))}}.$$

3.3.5.4 Intervalos de confianza para los cocientes de ventajas (odds-ratio).

En el caso del modelo de regresión logística múltiple se ha demostrado que los cocientes de ventajas de respuesta $Y = 1$ cuando se incrementa en una unidad cierta variable y se controlan las demás es de la forma

$$\theta(\Delta X_l = 1 | X_r = x_r, (r \neq l)) = e^{\beta_l} \quad \forall l = 1, \dots, R.$$

En este caso tomando exponenciales en el intervalo obtenido para cada uno de los parámetros se obtienen los siguientes intervalos de confianza aproximados al nivel de confianza $1 - \alpha$ para dichos cocientes de ventajas

$$e^{(\hat{\beta} \pm z_{\alpha/2} \hat{\sigma}(\hat{\beta}_r))}.$$

3.4 Modelos logit con variables explicativas categóricas.

En este apartado consideraremos el caso de variables explicativas categóricas que son aquellas cuyos valores no son susceptibles de medida sino que son un conjunto de cualidades exhaustivas y excluyentes. La metodología para construir el modelo de regresión logística con este tipo de variables consiste en asociarles variables cuantitativas que reciben el nombre de variables del diseño o variables ficticias y construir el modelo tomando como variables explicativas dichas variables del diseño. De este modo la estimación de parámetros y la inferencia sobre ellos se lleva a cabo mediante las técnicas tratadas hasta ahora para variables explicativas cuantitativas.

3.4.1 Variables del diseño.

Asociadas a una variable cualitativa A con categorías denotadas por $A_i (i = 1, \dots, I)$, se definen un total de $(I - 1)$ variables del diseño o variables ficticias. La razón para definir una variable ficticia menos que el número de categorías es que la matriz de diseño resultante del modelo de regresión logística múltiple correspondiente sea invertible (tenga columnas linealmente independientes).

A continuación se presentan distintas formas de codificación de las variables del diseño.

3.4.1.1 Método parcial.

Recibe también el nombre de codificación respecto a un grupo de referencia (reference cell coding). Consiste en elegir una categoría de referencia, de modo que todas las variables del diseño asignan el valor 0 a dicha categoría de referencia. Asociada a cada una de las restantes categorías se define una variable del diseño binaria que toma el valor 1 para su categoría asociada y el valor 0 para todas las demás. El programa estadístico SPSS da la posibilidad de elegir de entre la primera y la última categoría. Otros programas como SYSTAT y SAS toman como categoría de referencia la de mayor código (última).

De acuerdo con lo expuesto anteriormente, la m -ésima variable del diseño va asociada con la categoría A_m , y se define en la siguiente forma:

$$X_{im}^A = X_m^A(A = A_i) = \begin{cases} 1 & i = m \\ 0 & i \neq m \end{cases} \quad \forall m = 2, \dots, I; i = 1, \dots, I.$$

si es que se ha elegido la primera como categoría de referencia.

3.4.1.2 Método marginal.

Se llama también codificación mediante desviación respecto a la media (deviation from mean coding). De nuevo se toma una categoría de referencia a la que todas las variables del diseño asignan el mismo valor que en este caso es -1. Asociada a cada una de las restantes categorías se define una variable del diseño que toma el valor 1 para su categoría asociada y el valor 0 para todas las demás excepto para la de referencia (primera o última, según se haya elegido) a la que todas las variables ficticias asignan el valor -1.

Por lo tanto, la m -ésima variable del diseño está asociada con la categoría A_m y se define como

$$X_{im}^A = X_m^A(A = A_i) = \begin{cases} 1 & i = m \\ -1 & i = 1 \\ 0 & i \neq m, 1 \end{cases} \quad \forall m = 2, \dots, I; i = 1, \dots, I.$$

Observaciones

1. El tipo de codificación elegido dependerá de los objetivos perseguidos en el análisis estadístico.
2. El método usual es el de codificación con respecto a un grupo de referencia debido a que los parámetros tienen una interpretación sencilla en términos de cocientes de ventajas y suele ser de interés práctico estimar el riesgo de desarrollar una enfermedad para un grupo de exposición relativo a un grupo control no expuesto que se toma de referencia.
3. El método de desviación respecto a la media es el usado en Análisis de la Varianza para analizar la desviación de la media de la variable respuesta en cada nivel de la variable cualitativa y su media global. En el caso de regresión logística corresponderá al análisis de la desviación entre el logit en cada nivel de la variable cualitativa y la media de los logit.

3.4.1.3 Codificación de variables ordinales.

En el caso de una variable explicativa ordinal se puede proceder también de varias formas para construir el modelo de regresión logística

1. Se puede considerar como una variable nominal, definiendo a partir de ella las correspondientes variables del diseño con el método parcial o el marginal.
2. Codificar la variable ordinal asignando puntuaciones monótonas z_i a cada una de sus categorías $A_i (i = 1, \dots, I)$. La codificación puede ser arbitraria, asignando generalmente puntuaciones igualmente espaciadas, o basada en polinomios ortogonales.

Si se opta por la codificación de variables ordinales, éstas se introducirán en el modelo de regresión logística como variables cuantitativas cuyos valores son los códigos asignados a cada categoría.

3.5 Validación y diagnóstico del modelo logit.

Los estadísticos G^2 y χ^2 son medidas de la calidad global del ajuste. Una vez contrastado que un modelo se ajusta globalmente bien, se procede a estudiar mediante medidas alternativas la bondad del ajuste observación a observación, así como la naturaleza de la falta de ajuste.

Los métodos gráficos suelen ser de mucha ayuda. Otra forma habitual de validar un modelo es el estudio de los residuos que comparan el número observado de éxitos, en cada combinación de valores de las variables explicativas, con su valor ajustado por el modelo. En el caso de modelos GLM el estudio de los residuos puede poner de manifiesto si la falta de ajuste se debe a una elección inapropiada de la ligadura o a la falta de linealidad en los efectos de las variables explicativas.

También se pueden detectar observaciones influyentes calculando residuos y aproximando el efecto que produce sobre los parámetros borrar observaciones simples. En esta sección sólo consideraremos los métodos gráficos para validar el modelo, construidos a partir de la sensibilidad y especificidad.

3.5.1 Área bajo la curva ROC.

a) Sensibilidad, Especificidad y Valores predictivos.

Cuando se investiga la calidad de una prueba diagnóstica para determinada enfermedad, la sensibilidad se define como la proporción de enfermos correctamente identificados por los resultados positivos de la prueba y está dada por la relación $a/(a + c)$, siendo a el número de enfermos correctamente identificados y c el número de enfermos para los cuales la prueba de resultados negativos falsos.

La especificidad se define como la proporción de sanos correctamente identificados por los resultados negativos de una prueba y está dada por la relación $d/(b + d)$, siendo d el número de sanos correctamente identificados por la prueba y b el número de sanos para los cuales la prueba de resultados positivos falsos.

Otros de los componentes de la validez en las pruebas de detección son los valores predictivos de la prueba positiva y de la prueba negativa. La probabilidad de que un resultado positivo en la prueba indique un caso genuinamente enfermo, es el valor predictivo de la prueba positiva (VPPP). La probabilidad de que un resultado negativo sea genuinamente negativo, es el valor predictivo de la prueba negativa (VPPN).

Para determinar la calidad del examen diagnóstico debe probarse en sujetos de condición conocida: con o sin la enfermedad.

		Patología confirmada			
		Si	No		
Resultado del Procedimiento Diagnóstico	(+)	a Verdaderos	b Falsos	Sentido del Análisis ↓	
	(-)	c Falsos	d Verdaderos		
		(+)	(+)		
		(-)	(-)		

El análisis en sentido vertical de esta tabla da los siguientes sucesos posibles en términos de probabilidades condicionales:

1. Resultado (+) / enfermo = $\frac{a}{a+c}$ = *sensibilidad (S)*
2. Resultado (-) / sano = $\frac{d}{b+d}$ = *especificidad (E)*
3. Resultado (-) / enfermo = $\frac{c}{a+c}$ = *falso negativo*
4. Resultado (+) / sano = $\frac{b}{b+d}$ = *falso positivo*

b) Predicción de presencia o ausencia de enfermedad

Cuando un procedimiento diagnóstico se utiliza para detectar a los enfermos en una población, lo que interesa es el Valor Predictivo de la prueba Positiva (VPPP) y el Valor Predictivo de la Prueba Negativa (VPPN).

		Enfermos			
		Si	No		
Resultados del Procedimiento Diagnóstico	(+)	a	b	Sentido del Análisis →	
	(-)	c	d		
		(+)	(-)		
		(-)	(+)		

El análisis en sentido horizontal de esta tabla da los siguientes sucesos posibles:

1. Enfermo / Resultado (+) = $\frac{a}{a+b}$ = *VPPP*

2. Sano / Resultado (-) = $\frac{d}{c+d} = VPPN$
3. Sano / Resultado (+) = $\frac{b}{a+b} = \text{falso positivo de la predicción}$
4. Enfermo / Resultado (-) = $\frac{c}{c+d} = \text{falso negativo de la predicción}$

Cuando se trata de condiciones en que la falsa negatividad tiene consecuencias graves, por ejemplo que una persona tenga dengue hemorrágico y no tratarla, será importante preferir procedimientos diagnósticos de alta sensibilidad y después corroborarse el resultado con una prueba de alta especificidad.

El valor predictivo del resultado de la prueba depende no sólo de la calidad de la prueba sino también de la prevalencia de la enfermedad en la población que se define como:

$$\frac{\text{Número de enfermos}}{\text{Número de personas en la población}}$$

Es importante definir lo anterior porque la prevalencia no afecta la sensibilidad y la especificidad de una prueba, sí influye en el valor predictivo del resultado del examen.

c) Aplicación del teorema de Bayes

Simolicemos por “R” un resultado positivo del examen y por “E” el hecho de tener la enfermedad. Conociendo la sensibilidad y la especificidad del examen diagnóstico, el VPPP, sería la probabilidad de tener la enfermedad dado que el resultado es positivo: $P(E|R)$ se expresa como:

$$P(E|R) = \frac{P(E)P(R|E)}{P(E)P(R|E) + P(\bar{E})P(R|\bar{E})}$$

Donde:

$P(E)$ Equivale a la prevalencia de la enfermedad en la población,

$P(R|E)$ Es la sensibilidad y

$P(R|\bar{E})$ El complemento de la especificidad.

Es decir:

$$VPPP = \frac{(\text{Prevalencia})(\text{Sensibilidad})}{(\text{Prevalencia})(\text{Sensibilidad}) + (1 - \text{Prevalencia})(1 - \text{Especificidad})}$$

De igual forma el valor predictivo de un examen negativo será:

$$P(\bar{E}|\bar{R}) = \frac{P(\bar{E})P(\bar{R}|\bar{E})}{P(\bar{E})P(\bar{R}|\bar{E}) + P(E)P(\bar{R}|E)}$$

Lo que en términos epidemiológicos sería:

$$VPPN = \frac{(1 - \text{Prevalencia})(\text{Especificidad})}{(1 - \text{Prevalencia})(\text{Especificidad}) + (\text{Prevalencia})(1 - \text{Sensibilidad})}$$

El área bajo la curva ROC (Receiver Operating Characteristic), cuyos rangos están entre 0 y 1, es una medida que muestra la habilidad del modelo para discriminar a los individuos enfermos y los no enfermos.

Como regla general, si $ROC = 0.5$, no hay discriminación, si $0.7 \leq ROC < 0.8$, discriminación aceptable y si $0.8 \leq ROC < 0.9$ la discriminación es excelente.

4. RESULTADOS Y DISCUSIÓN.

4.1 Prevalencia de factores de riesgo.

En el presente trabajo no se incluyen las variables numéricas tales como peso, estatura, medida de la cintura, presión sistólica y glucemia, porque a partir de estas se han construido algunas de las variables categóricas incluidas. Dado que los datos se generaron originalmente para un estudio de obesidad y diabetes, no se incluyó la condición de fumador del paciente y las variables consideradas como contribuyentes.

La tabla 2 muestra la prevalencia de cada factor de riesgo cardiovascular en la población estudiada. Se encontró que 10.5% son hipertensos, a diferencia de la investigación realizada por (Campos-Nonato, Hernández-Barrera, Rojas-Martínez, Pedroza-Tobías, Medina-García, & Barquera, 2013) en la que encontraron que el 31.5% de los adultos mexicanos de 20 años a más padecen hipertensión, el 12.2% obtuvo

niveles altos de colesterol LDL, se encontró un 94.2% con bajos niveles de colesterol bueno (HDL), el 17.4% presentaron un cuadro clínico de diabetes, el 40.7% de la población son obesos, el 36.6% se encontró en sobrepeso, también se encontró que el 66.3% no realiza ejercicio al menos 30 minutos, el 34.3% son hombres, el 18.6% son mayores de 55 años y el 28% tiene antecedentes familiares con diabetes

Tabla 2: Prevalencia de factores de riesgo cardiovascular.

Factores de Riesgo Cardiovascular	%	Bootstrap para prevalencia		
		E. Típ.	IC al 95%	
			Inf.	Sup.
Hipertensión	10.5	2.4	5.8	15.1
Colesterol LDL	12.2	2.5	7.6	17.4
Colesterol HDL(Menor de 35 mg/dl)	94.2	1.7	90.7	97.1
Diabetes	17.4	2.7	12.2	23.2
Obesidad	40.7	3.8	33.1	48.3
Sobrepeso	36.6	3.7	29.7	44.2
Ejercicio al menos 30 minutos	66.3	3.5	59.9	73.3
Mayor de 55 años de edad	18.6	3.0	12.8	25.0
Antecedentes familiares con diabetes	27.9	3.4	21.5	34.3

Fuente: Elaboración propia.

De acuerdo a la definición de riesgo cardiovascular, un paciente se encuentra en riesgo cuando presenta al menos uno de los factores de riesgo mencionado, por lo que, como se observa en la tabla 2, aunque es evidente el RC debido a que 94.2% los pacientes presentan concentraciones bajas de colesterol bueno (HDL) y el 66% no hace ejercicio por al menos 30 minutos, con estos elementos no es suficiente para afirmar que la población tiene alta probabilidad de un evento cardiovascular. Otra variable que tiene prevalencia alta es la obesidad y sobrepeso con 4 de cada diez pacientes, aproximadamente.

Uno de los principales factores de riesgo cardiovascular es el llamado colesterol malo o colesterol LDL, entonces, considerando la población global se encuentra que están riesgo cardiovascular el 12.2% de lapoblación estudiada, por su alta concentración de LDL, como se observa en la tabla 3.

Tabla 3: Prevalencia de RC según LDL en pacientes de la muestra.

Riesgo Cardiovascular	%	Bootstrap para prevalencia de RC		
		E. Típ.	IC al 95%	
			Inf.	Sup.
Alta concentración de LDL(Si)	12.2	2.5	7.6	17.4
Alta concentración de LDL(No)	87.8	2.5	82.6	92.4

Fuente: Elaboración propia.

Se investigó la relación que existe entre RC y los factores de riesgo cardiovascular a partir de una prueba de independencia en tablas de contingencia. La prueba de independencia chi-cuadrado mostró evidencia de que todos los factores de riesgo estudiados contribuyen en alguna medida para incrementar el riesgo cardiovascular, pero los que tienen mayor contribución son niveles altos de LDL, niveles bajos de HDL, no hacer ejercicio por lo menos 30 minutos al día y antecedentes familiares con diabetes. La tabla 4 confirma esta afirmación y se observa que el $OR = 0.068$ (menor de 35mg/dl)/(mayor de 35mg/dl) o de forma equivalente usando su recíproco, $OR = 1/0.068 = 14.7$ (mayor de 35mg/dl)/ (menor de 35mg/dl), nos indica que el riesgo cardiovascular en las personas con niveles de colesterol inferiores a 35mg/dl, es superior hasta en 14 veces que en personas que reportan niveles superiores a los 35mg/dl.

Tabla 4: Comparación RC vs Factores de RC

Comparación		OR	IC al 95%		p-Valor
RC	Razón de las ventajas para Hipertensión (Si / No)	2.303	.680	7.800	.170
	Colesterol LDL (Altas concentraciones/Bajas concentraciones)	-	-	-	0.00
	Razón de las ventajas para Colesterol HDL (Menor de 35 mg/dl. Riesgo Cardiovascular / Mayor de 35 mg/dl. Menor riesgo Cardiovascular.)	.068	.017	.268	0.00
	Razón de las ventajas para Diabetes (Si / No)	2.117	.746	6.003	.151
	Razón de las ventajas para Obesidad (Si / No)	2.138	.848	5.388	.102
	Razón de las ventajas para ¿Hace ejercicio al menos 30 minutos? (Si / No)	3.828	1.484	9.870	.004
	Razón de las ventajas para Sexo (Femenino / Masculino)	.829	.323	2.128	.696
	Razón de las ventajas para Mayor de 55 años de edad (No / Sí)	.967	.302	3.098	.956
	Razón de las ventajas para Antecedentes familiares con diabetes (Si / No)	2.703	1.065	6.865	.032

Fuente: Elaboración propia.

El riesgo cardiovascular en pacientes que no hacen ejercicio por lo menos 30 minutos es hasta 4 veces superior que en pacientes que se ejercitan por lo menos 30 minutos. Observe el OR=3.83 Si/No.

4.2 Puntuaciones Framingham para la población.

Usando las fórmulas del modelo de Framingham, se calculan las probabilidades de riesgo cardiovascular y encontramos que la probabilidad promedio de riesgo cardiovascular para toda la población es de 8.62% y un intervalo de confianza (7.32, 9.97) al 95%, obtenido mediante bootstrap. El criterio clásico para la clasificación de RC es que el paciente tenga una puntuación de 0.2 o mayor, es decir, la probabilidad debe ser mayor o igual al 20% de la puntuación de Framingham. Con base a este criterio se clasifican como pacientes en RC el 2, 14, 26, 41, 49, 62, 67, 70, 75, 79, 84, 97, 110, 114, 130, 156, 157 y 169, en total 18 personas.

Tabla 5: Probabilidad promedio de RC con el modelo de Framingham.

	Statistic	Bootstrap ^a		
		Típ. Error	Intervalo de confianza al 95%	
			Inferior	Superior
Media	.08621428	.00671155	.07273643	.09943690
Desv. típ.	.089821173	.007613955	.074068992	.104428827

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Fuente: Elaboración propia.

Las probabilidades de riesgo cardiovascular a los 10 años, para los primeros diez pacientes se indican en la tabla 6, en la que se observa que se encuentran en riesgo cardiovascular **el paciente 2, aunque habrá que vigilar a los pacientes 3 y 5.**

Tabla 6: Tabla de probabilidades de riesgo cardiovascular para los primeros 10 pacientes con el modelo de Framingham.

Paciente	1	2	3	4	5	6	7	8	9	10
Probabilidad	0.074	0.4118	0.1239	0.074	0.147	0.0275	0.0797	0.0242	0.0354	0.035375

Fuente: Elaboración propia.

Los cálculos para los pacientes 1 y 2, que son mujer y hombre, se han realizado de la forma indicada en la siguiente tabla:

Tabla 7: Ejemplo numérico sobre RC con las puntuaciones de Framingham

Caso Femenino		Caso Masculino	
$E_1 * \text{Edad} = 0.33766 * 49$	16.545	$E_1 * \text{Edad} = 0.04826 * 69$	0.048
$E_2 * \text{Edad}^2 = -0.00268 * 49$	-6.435		0.000
b_c de 200-239 le corresponde	0.208	b_c de 200-239 le corresponde	0.177
$b_{HDL} < \text{que } 35$ le corresponde	0.843	$b_{HDL} < \text{que } 35$ le corresponde	0.497
b_t PAS < 120, PAD < 80 corresponde	-0.534	b_t PAS < 140, PAD < 90 corresponde	0.283
b_D No tiene diabetes	0.000	b_D Tiene diabetes	0.428
$L_M = b_{E1} * \text{Edad} + b_{E2} * \text{Edad}^2 + b_C + b_H + b_T + b_D$	10.628	$L_H = b_{E1} * \text{Edad} + b_C + b_H + b_T + b_D$	4.716
G_M	9.925	G_H	3.098
$B = \exp(L_M - G_M)$	2.019	$B = \exp(L_H - G_H)$	5.045
$Y_{mujer} = 1 - S^B = 1 - 0.96246^{2.018611704}$	0.074	$Y_{hombre} = 1 - S^B = 1 - 0.90015^{5.044961387}$	0.412

Fuente: Elaboración propia.

Continuando con la idea de clasificación de los pacientes en alto riesgo cardiovascular, proponemos enseguida una clasificación por regresión logística, tomando como variable de estado el RC a partir de la puntuación de Framingham. Designemos a esta variable *Riesgo cardiovascular según puntuación de Framingham*, cuyos valores serán Sí y No.

4.2.1 Modelo de regresión logística.

El ajuste de un modelo para estimar la probabilidad de que un paciente clasifique como en alto riesgo cardiovascular, dados algunos factores de riesgo, es factible con regresión logística.

Para construir el modelo de regresión logística, se incluyen como factores de riesgo las variables hipertensión = T, colesterol elevado o lipoproteína de baja densidad = C, lipoproteína de alta densidad = H, diabetes mellitus tipo II = D, obesidad = OB, inactividad física = IF, sexo = (Hombre, Mujer), edad = E y antecedentes familiares con diabetes = AFCD, consideradas estadísticamente significativas. La variable de interés es, por supuesto, alto riesgo cardiovascular (RCF), con las categorías Si = 1 y No = 0. Se ejecuta el procedimiento en software estadístico, utilizando el método de selección *stepwise* de Wald, un punto de corte de 0.2 y se obtienen los resultados de bondad de ajuste global dados en las tablas 8, 9 y 10.

Como se puede observar en la tabla 8, la prueba de razón de verosimilitud (-2log de verosimilitud) da un valor de 47.403 en el último paso del método, lo cual indica que el modelo es aceptable, también el R cuadrado de Nagelkerke indica que el modelo explica un 67% de la variabilidad.

Tabla 8: Resumen del modelo.

Paso	-2 log de la verosimilitud	R cuadrado de Cox y Snell	R cuadrado de Nagelkerke
1	77.898 ^a	.195	.400
2	51.920 ^b	.308	.631
3	47.403 ^b	.326	.668

Fuente: Elaboración propia.

Otro elemento que proporciona información sobre la bondad del modelo es la prueba de Hosmer y Lemeshow, que en el último paso del método se observa un p-valor 0.074, superior al valor de prueba p-valor = 0.05.

Tabla 9: Prueba de Hosmer y Lemeshow

Paso	Chi cuadrado	gl	Sig.
1	.000	0	.
2	1.407	2	.495
3	6.947	3	.074

Fuente: Elaboración propia.

Un recurso adicional para probar la bondad del modelo, es la tabla de clasificaciones correctas, que en este caso se observa una clasificación global de 91.3%, mientras que la clasificación para los pacientes en

bajo riesgo cardiovascular es del 92.2% y aquellos que se encuentran en alto riesgo cardiovascular representan el 83.3%.

Tabla 10: Tabla de clasificación^a

Riesgo Cardiovascular según puntuación de Framingham	No	92.2
	Si	83.3
Porcentaje global		91.3

Fuente: Elaboración propia.

4.2.1.1 Estimación de los parámetros.

Se estiman los parámetros por MV y se realiza la prueba de hipótesis sobre la igualdad a cero de los parámetros.

Tabla 11: Variables en la ecuación.

Variables	B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
							Inferior	Superior
Diabetes(1) = D	3.881	.917	17.902	1	.000	48.472	8.030	292.595
Sexo(1)	-1.633	.809	4.076	1	.043	.195	.040	.953
EdadMayor55(1) = E	-3.720	.913	16.599	1	.000	.024	.004	.145
Constante	-.811	.663	1.500	1	.221	.444		

Fuente: Elaboración propia.

Encontramos que son significativamente distintos de cero los coeficientes de D, Sexo y E. Entonces el modelo de regresión logística se expresará como

$$P[RC|X] = \frac{1}{1 + e^{-(-0.811 + 3.881D - 1.633Sexo - 3.72E)}}$$

a partir del cual podemos estimar las probabilidades de clasificar en el grupo de alto riesgo cardiovascular a los pacientes. La tabla 13 representa las probabilidades de los primeros diez pacientes y a manera de ejemplo, realizamos los cálculos para los pacientes 1 y 2, considerando mujer=1 y hombre=0; Diabetes (Si=1, No=0); E=Adad>55años (1=No, 0=1). A continuación el ejemplo de los primeros dos sujetos:

Sexo	Diabetes	Edad
Femenino (1)	No (0)	49 (0)
Masculino (0)	Si (1)	69 (1)

Tabla 12: Ejemplo numérico sobre RC con las probabilidades del modelo logit.

Caso Femenino	Coso Masculino
Para D=0, Sexo=Mujer=1 y E>55=1	Para D=1, Sexo=Hombre=0 y E>55=0
$P[RC D, Sexo, E]$	$P[RC D, Sexo, E]$
$= \frac{1}{1 + e^{-(-0.811 + 3.881(0) - 1.633(1) - 3.72(1))}}$	$= \frac{1}{1 + e^{-(-0.811 + 3.881(1) - 1.633(0) - 3.72(0))}}$
= 0.0021	= 0.95562

Fuente: Elaboración propia.

Entonces, los cálculos corresponden a una mujer, sin diabetes y con una edad no mayor a los 55 años. Para el caso del hombre, tiene diabetes y su edad es mayor que los 55 años.

Tabla 13: Tabla de probabilidades de riesgo cardiovascular para los primeros 10 pacientes con el modelo de regresión logística, usando el tercer cuartil de las puntuaciones de Framingham.

Paciente	1	2	3	4	5	6	7	8	9	10
Probabilidad	0.002	0.9556	0.002	0.002	0.3429	0.002	0.002	0.002	0.0106	0.002
d	1	2	1	1	7	1	1	1	5	1

Fuente: Elaboración propia.

Como se observa de los datos, los pacientes 2 y 5 clasifican en riesgo cardiovascular con probabilidad superior al punto de corte 0.2, como ocurre con el modelo de Framingham, aunque aquí el

paciente 3 tiene un valor de probabilidad cercano a cero, lo cual es evidencia que no se encuentra en riesgo cardiovascular. La recomendación sería atender de inmediato a los pacientes 2 y 5, aunque el de mayor riesgo es el paciente 2.

5. CONCLUSIONES.

Diagnosticar a un paciente en riesgo cardiovascular no debe resultar difícil si atendemos la definición de RC, la dificultad médica radica en identificar los factores de riesgo que concurren en un paciente, ¿Cuál debe ser el tratamiento médico apropiado para reducir la probabilidad de un evento cardiovascular? En otras palabras, ¿Cómo controlamos simultáneamente en un paciente todos los factores de RC? Los especialistas coinciden en que esto podría hacerse promoviendo entre la gente la actividad física y mejores prácticas alimentarias, razón por la cual dentro de las jornadas de la cruzada, se contó con preparadores físicos y un equipo de nutriólogas, que a todos los pacientes les establecieron una rutina de actividad física (ejercicios aeróbicos) y recomendaciones prácticas en su dieta y alimentación.

Todos los pacientes con factores de riesgo y con diagnóstico de obesidad, diabetes e HTA recibieron atención médica y fueron canalizados a los servicios de salud más cercano, dentro de su jurisdicción, para su seguimiento y tratamiento adecuado.

En modelo de Framingham, a partir de las puntuaciones, resultó de utilidad en la clasificación de pacientes en alto riesgo cardiovascular, aunque el modelo de regresión logística propuesto permite discriminar, con una tasa de 91.3% de clasificaciones correctas, entre pacientes en RC y No RC con solo las variables *diabetes (D)*, Sexo (Hombre, Mujer) y edad mayor de 55 años, que en términos de costos, es mejor que el modelo de Framingham, toda vez que para el diagnóstico utiliza un número menor de variables.

Con regresión logística valores de probabilidad cercanos a 1 indican alto riesgo cardiovascular y el paciente debe ser atendido de inmediato, mientras que valores cercanos a 0 indican baja probabilidad de riesgo cardiovascular y los factores de riesgo presentes en el paciente pueden tratarse de forma ordinaria. Los cálculos de las probabilidades de pueden hacer en una calculadora de bolsillo o en un libro de cálculo tipo Excel teniendo en cuenta que 1 punto de corte para este caso es de 0.2

Por tanto se concluye que el modelo de Framingham es bueno para la clasificación de pacientes en alto riesgo cardiovascular, pero también lo es el modelo de regresión logística propuesto, por lo que es posible utilizar cualquiera de ellos para estimar la probabilidad de que un paciente sufra un evento cardiovascular en un periodo de 10 años, a partir de la observación de pocas variables.

REFERENCIAS

- [1] Agresti, A. (1996). An Introduction to Categorical Data Analysis. New York: JOHN WILEY & SONS, INC.
- [2] Alcocer, L. A., Lozada, O., Fanghanel, G., Sánchez-Reyes, L., & Campos-Franco, E. (2011). Estratificación del riesgo cardiovascular global. Comparación de los métodos Framingham y SCORE en población mexicana del estudio PRIT. Cirugía y Cirujanos, Vol. 79, No. 2, pp. 168-174.
- [3] Alejandro, S. J. (9 de Octubre de 2008O). Chocolate bueno para el corazón. Recuperado el 2 de Julio de 2016, de Así sucede: <http://asisucedo.com.mx/chocolate-bueno-para-el-corazon/>
- [4] Campos-Nonato, I., Hernández-Barrera, L., Rojas-Martínez, R., Pedroza-Tobías, A., Medina-García, C., & Barquera, S. (2013). Hipertensión arterial: prevalencia, diagnóstico oportuno, control y tendencias en adultos mexicanos. Salud Pública Méx 2013, Vol. 55(sup 2):144-150.
- [5] DMEDICINA.com, Salud y bienestar. (9 de Octubre de 2015). Factores de riesgo cardiovascular. Recuperado el 2 de Julio de 2016, de <http://www.dmedicina.com/enfermedades/enfermedades-vasculares-y-del-corazon/factores-de-riesgo-cardiovascular.html>.
- [6] ENSANUT. (2012). Módulo del Programa de Oportunidades en la Encuesta Nacional de Salud y Nutrición. México: Instituto Nacional de salud Pública.
- [7] Farmacias ABC. (2006). Riesgo cardiovascular, la importancia de la prevención. Recuperado el 06 de Julio de 2016, de <http://www.farmacias-abc.com.ar/contenido/Index624.asp>.
- [8] Gobierno Federal, SALUD, SEDENA Y SEMAR. (s.f.). Detección y Estratificación de Factores de Riesgo Cardiovascular. Recuperado el 02 de Julio de 2016, de

- http://www.cenetec.salud.gob.mx/descargas/gpc/CatalogoMaestro/421IMSS_421_11_Factores_riesgo_cardiovascular/GRR_FACTORES_RIESGO_CARDIOVASCULAR.pdf.
- [9] Gueyffier, F., Boissel, J. P., Bouitief, F., Pocock, S., Coope, J., Cutler, J., y otros. (1997). Effect of antihypertensive treatment in patients having already suffered from stroke. Gathering the evidence. The INDANA (INdividual Data ANalysis of Antihypertensive intervention trials) Project Collaborators. *Stroke*, Dec, 28 (12): 2557-2562.
- [10] Hosmer, D. W., & Lemeshow, S. (1999). *Applied Survival analysis. Regression Modeling of Time to Event Data*. New York: JOHN WILEY & SONS, INC.
- [11] Iglesias, T. (2013). *Métodos de Bondad de Ajuste en Regresión Logística*. Tesis. Granada, España: Universidad de Granada.
- [12] López, J. (9 de Diciembre de 2013). *Enfermedades cardiovasculares*. Recuperado el 2 de Julio de 2016, de Prezi: <https://prezi.com/y8f6izd9jnks/enfermedades-cardiovasculares/>
- [13] Molinero, L. M. (2003). *Modelos de riesgo cardiovascular. Estudio de Framingham. Proyecto SCORE*. Asociación de la Sociedad Española de Hipertensión. Liga Española para la Lucha contra la Hipertensión Arterial.
- [14] OMS. (Enero de 2015). *Enfermedades cardiovasculares*. Recuperado el 2 de Julio de 2016, de Organización Mundial de la Salud: <http://www.who.int/mediacentre/factsheets/fs317/es/>
- [15] QUO. (2 de mayo de 2012). *El 30% de mexicanos son hipertensos*. Recuperado el 2 de Julio de 2016, de <http://quo.mx/noticias/2012/05/02/el-30-de-mexicanos-son-hipertensos>
- [16] Rodríguez, P. (2014). Tesis. Oviedo, Asturias, España: Universidad de Oviedo.
- [17] Ruiz, E., Segura, L., & Agusti, R. (2012). *Uso del Score de Framingham como Indicador de los Factores de Riesgo de las Enfermedades Cardiovasculares en la población Peruana*. *Revista Peruana de Cardiología*, Vol. XXXVII, N° 3.
- [18] Texas Heart Institute. (Octubre de 2015). *Factores de riesgo cardiovascular*. Recuperado el 6 de Julio de 2016, de http://www.texasheart.org/HIC/Topics_Esp/HSmart/riskspan.cfm
- [19] WHO. (Octubre de 2003). *The World Health Report 2003*. Recuperado el 2 de Julio de 2016, de World Health Organization: http://www.who.int/whr/2003/en/whr03_en.pdf?ua=1

Capítulo 9

pp 109-116

CIFRADO DE IMÁGENES MÉDICAS USANDO OPERACIONES DE ADN Y MAPAS CAÓTICOS

Luis René Marcial Castillo, Marcela Rivera Martínez, Lourdes Sandoval Solís, Aldo Rafael Matus Angulo, Erika Leonor Basurto Munguía
Benemérita Universidad Autónoma de Puebla
Puebla, Puebla, México

ABSTRACT

In this paper, it was designed and computationally implemented an algorithm for symmetric cryptography which is based on chaotic maps and operations of addition and subtraction of DNA. The cryptographic system were implemented in the OCTAVE programming language, it allows encrypting and decrypting medical images of any size. The tests show that the computational implementation encrypts and decrypts correctly.

KEYWORDS: Medical imaging, symmetric cryptography, chaotic maps, deoxyribonucleic acid

RESUMEN

En este trabajo, se diseñó e implementó computacionalmente un algoritmo para criptografía simétrica el cual se basa en mapas caóticos y operaciones de adición y sustracción del ADN. El sistema criptográfico implementado en el lenguaje de programación OCTAVE permite cifrar y descifrar imágenes médicas de cualquier dimensión. Las pruebas realizadas muestran que la implementación computacional cifra y descifra de forma correcta.

1. INTRODUCCIÓN

Criptografía

La confidencialidad de las comunicaciones es de suma importancia en la sociedad moderna. La industria, gobierno y particulares confían en que la tecnología les garantice que el intercambio de datos sea seguro de modo que no se permita a terceros acceder al contenido de tal comunicación, la cuestión de la confidencialidad ha sido dejada al campo de la criptografía [3]. Goldreich transcribe el planteamiento de la criptografía como "el problema de proveer comunicación secreta sobre medios inseguros" [8]. La criptografía es una herramienta muy útil cuando se desea tener seguridad informática, es decir, cuando se cuenta con un medio para garantizar las propiedades de confidencialidad, para lograrlo, se crean mecanismos que garanticen en cierta medida las propiedades de disponibilidad, integridad y confidencialidad. La disponibilidad se refiere a que la información siempre este presente, la integridad es que no se pierda información, la confidencialidad se puede lograr usando mecanismos que aunque sea robada la información, permita que no se pueda acceder a esta o garantice de alguna forma que no se pueda llegar a ella, hasta que pierda su valor. Estos mecanismos permiten ver si la información ya creada ha sufrido o no alguna modificación no autorizada. El criptoanálisis, también llamado "criptología" es la disciplina contraria a la criptografía, se encarga de analizar la información cifrada para revelar la información original sin necesidad de las claves secretas y de esta forma romper los procedimientos previamente establecidos por la criptografía, el criptoanálisis lo usan los investigadores como una forma de probar las fortalezas o debilidades de los cripto sistemas [6].

La criptografía se divide en dos grandes ramas, la criptografía de clave privada o simétrica y la criptografía de clave pública o asimétrica. La primera se refiere al conjunto de métodos que permiten una comunicación segura entre las partes siempre que con anterioridad se intercambie la clave correspondiente que se denomina clave o llave simétrica. La criptografía simétrica utiliza algún método matemático para cifrar o descifrar un mensaje utilizando únicamente la llave secreta, esta llave se debe compartir con las personas que se desea vean los mensajes. El problema de la criptografía simétrica es compartir la clave de forma confidencial, pero esto se resuelve con el uso de la criptografía asimétrica. Uno de los sistemas criptográficos simétricos más conocidos es el algoritmo DES (Data Encryption Standard), la U.S. National Institute of Standards and Technology (NIST) lo estandarizó en los años 70, el mejor ataque conocido al algoritmo DES es la búsqueda exhaustiva de la llave, los especialistas en cómputo diseñaron una supercomputadora usando las PCs conectadas a la red y usando el poder de 100000 PCs fueron capaces de encontrar la llave usada por el algoritmo DES después de 22 horas y 15 minutos de usar la supercomputadora, después de este resultado, la

opción que se tomó fue sustituir el algoritmo DES por el cifrado múltiple, que consiste en aplicar varias veces el mismo algoritmo con el fin de fortalecer la longitud de la clave, esto ha tomado forma como un nuevo sistema para el cifrado, conocido actualmente como el algoritmo TDES (Three Data Encryption Standard), aunque actualmente el estándar de la criptografía simétrica es el algoritmo AES (Advanced Encryption Standard), el cual fue adoptado desde 1997 por la U.S. NIST y el algoritmo usado es el propuesto por Rijndael en el año 2002 [3].

La criptografía de clave pública o asimétrica, también es por definición aquella que utiliza dos llaves o claves diferentes para cada usuario: una llave se usa para cifrar y la otra se usa para descifrar (conocida como clave o llave privada). Lo que se cifra con una llave se puede descifrar con la otra llave, si n personas requieren enviar un mensaje cifrado, sólo es necesario que los demás personas conozcan su llave pública, esto reduce el número de llaves, ahora el problema está en que la llave pública sea de la persona que dice ser (autenticación de las llaves públicas). El estándar de la criptografía asimétrica es el algoritmo RSA (en honor a sus inventores Rivest, Shamir y Adleman) publicado en 1978 [4].

Ácido desoxirribonucleico

El ácido desoxirribonucleico (ADN) es el material genético de todos los organismos celulares y de casi todos los virus. El ADN lleva la información necesaria para dirigir la síntesis de proteínas y la replicación. Cada molécula de ADN está constituida por dos cadenas o bandas formadas por un elevado número de compuestos químicos llamados nucleótidos. Estas cadenas forman una especie de escalera retorcida que se llama doble hélice. Cada nucleótido está formado por tres unidades: una molécula de azúcar llamada desoxirribosa, un grupo fosfato y uno de cuatro posibles compuestos nitrogenados llamados bases, las cuales son: Adenina (A), Guanina (G), Tiamina (T) y Citosina (C). Conforme al complemento presentado por Watson y Crick, la Adenina es complementada por Tiamina y Guanina con Citosina [1]. En este método la Adenina se codifica con 00, la Tiamina con 11, la Guanina con 10 y la Citosina con 01 [16]. La regla de adición (suma binaria) para el ADN es realizada como lo muestra el cuadro 1, por ejemplo, si se desea realizar la operación T+T el resultado sería G, ya que la suma binaria $11+11$ es igual a 110 pero al desechar el bit más significativo (más a la izquierda) se obtiene el número binario 10 que corresponde a G. La regla de sustracción (resta binaria en complemento a 2) es mostrada en el cuadro 2, por ejemplo, el resultado de la operación T-T es A, ya que el complemento a 2 de T es: $11 \rightarrow 00$ (se aplica complemento a 1, es decir, se cambia 1 por 0 y 0 por 1) $\rightarrow 00+1$ (se suma 1 al bit menos significativo) y el resultado es 01, al sumar $T+01$ resulta que $11+01=100$, en donde se desecha el bit más significativo y el resultado es 00 que corresponde a la letra A.

Cuadro 1. Regla de adición para el ADN.

+	A	C	G	T
A	A	C	G	T
C	C	G	T	A
G	G	T	A	C
T	T	A	C	G

Cuadro 2. Regla de sustracción para el ADN.

-	A	C	G	T
A	A	T	G	C
C	C	A	T	G
G	G	C	A	T
T	T	G	C	A

Métodos caóticos

Puede decirse que la dinámica caótica inició con el trabajo del matemático Francés Henri Poincaré a finales del siglo XIX. La motivación de Poincaré fue promovida por el problema de las orbitas de tres cuerpos celestes experimentando atracción gravitacional mutua (por ejemplo, una estrella y dos planetas). Poincaré fué capaz de mostrar que orbitas muy complicadas eran posibles (ahora llamadas caóticas). No obstante, la posibilidad de caos en sistemas físicos reales no fue ampliamente apreciada sino hasta la actualidad, mucho del crédito por este cambio es atribuido a la extensa solución numérica de sistemas dinámicos en computadoras digitales [12]. En los últimos años, los hilos del caos y la dinámica no lineal se han esparcido a través de disciplinas científicas como una intrincada red araña. Caos y dinámica no lineal han provisto de nuevas herramientas teóricas y conceptuales que permiten capturar, entender y enlazar los compartimientos complejos de sistemas simples -el tipo de comportamiento llamado caos en la ciencia contemporánea- [9]. Se puede decir, que una función caótica es una función matemática que describe un sistema dinámico no lineal complejo cuya evolución en el tiempo hace imposible la predicción a largo plazo, luciendo errático y casi aleatorio. Los ejemplos más notorios de las características del caos son el llamado *efecto mariposa* y la *impredecibilidad de órbitas pseudo-aleatorias*, generadas por ecuaciones deterministas. Estos fenómenos, así como otros relacionados con el caos han sido tradicionalmente asociados a mecanismos de *confusión* y *difusión* los cuales son la base principal de un buen sistema criptográfico.

Propuesta de este trabajo

Varios sistemas de protección se basan en los estándares de la criptografía simétrica como lo son los algoritmos TDES y AES, los cuales tienen la característica de ser cripto sistemas muy seguros. La desventaja que tienen estos métodos es que en su proceso de cifrado y descifrado son muy lentos y esto es un gran problema cuando se tienen grandes volúmenes de datos como sucede en el caso de las imágenes digitales. Existen diversas disciplinas que requieren cifrar de forma rápida y enviar los datos cifrados como los servicios de telemedicina, por citar alguno.

Las funciones caóticas desarrolladas como modelos matemáticos de sistemas no lineales son completamente sensibles a cambios muy simples en las condiciones iniciales, esto les proporciona gran aplicabilidad en diversos campos, como lo es la criptografía.

La propuesta de este trabajo, es desarrollar la implementación computacional bajo el ambiente del lenguaje de programación OCTAVE de un cripto sistema que use ventajas mencionadas de las funciones caóticas, la función caótica se usa para generar la llave del proceso de cifrado y descifrado, posteriormente se combina la llave con las operaciones de ADN mencionadas en los cuadros 1 y 2 con el objetivo de mejorar la seguridad del sistema criptográfico simétrico que se implementa. La finalidad de usar este esquema es cifrar y descifrar de forma rápida imágenes digitales sin la lentitud del estándar AES de la criptografía simétrica.

El trabajo se organiza de la manera siguiente: la sección 2, presenta los antecedentes del proyecto en lo que se refiere al uso de funciones caóticas y operaciones bajo ADN, en la sección 3, se muestran los pasos de los algoritmos que realizan el cifrado y descifrado, la sección 4 proporciona las pruebas del cripto sistema usando imágenes médicas sobre las enfermedades de: rubeola, sífilis secundaria, tiroides y ojos. En la sección 5 se presentan las conclusiones del trabajo y finalmente se listan las referencias.

2. ANTECEDENTES DEL PROYECTO

Desde la década de los 90's muchos investigadores han notado que existe una importante relación entre el caos y la criptografía: muchas propiedades de los sistemas caóticos tienen sus correspondientes contrapartes en los cripto sistemas tradicionales. Los sistemas caóticos pueden conocer sus ecuaciones y sus condiciones iniciales fijas, sin embargo, la más mínima variación provoca una evolución radical en su comportamiento [3].

La teoría de funciones y mapas caóticos se presentan en varios trabajos. Fuan, Mengb, Zhanb, Zhuc, Laud, Tsed y Mae en el 2013 [5] proponen un esquema de protección de imágenes médicas basado en mapas caóticos; Gao y Chen en el 2008 [7] proponen un algoritmo nuevo de permutación de píxeles; Huang y Nien en el 2009 [10] proponen un sistema multi-caótico basado en el mismo principio de permutación de Gao y Chen; Patidar, Pareek y Sud en el 2009 [13] proponen un cifrado tipo sustitución-difusión basado en mapas logísticos y caóticos; Rhouma, Meherzi y Belghith en el 2009 [14] proponen el cifrado de imágenes a color basado en mapas caóticos; Sun, Liu y Li en el 2008 [15] proponen un esquema de encriptación basado en

mapas caóticos espaciales; Tong y Cui en el 2009 [17] proponen un generador de cifrado de secuencias caóticas con componentes dinámicos; Wong, Kwok y Law en el 2008 [19] proponen un esquema de encriptación basado en el mapa caótico estándar; Xiao y Xia en el 2009 [20] proponen un esquema de encriptación usando mapas de permutación; Xu, Wang y Yang en el 2008 [21] proponen una mejora en el algoritmo de encriptación de imágenes que usan mapas caóticos y Ye en el 2009 [22] propone un cripto-sistema basado en las matrices de Töplitz y Hankel.

En lo que se refiere a la teoría de secuencias de ADN se presenta en varios trabajos. Tere, Vaida, Alboai y Chiorean, en el 2011 [16] proponen el uso de ADN para criptografía simétrica; Anwarl, Paul y Singh en el 2014 [2] hacen una revisión de la forma en que se ha realizado la transmisión de mensajes basada en el uso de ADN; Javheri y Kulkarni en el 2014 [11] proponen también un algoritmo criptográfico para la comunicación segura de datos basado en ADN; Anil y Chirakkarottu en el 2014 [1] presentan un método de encriptación para el iris del ojo humano basado también en operaciones de ADN.

Como lo muestran los trabajos mencionados las funciones caóticas han llamado la atención de varios investigadores, razón por la cual en este trabajo se plantea su uso bajo la combinación de las operaciones de adición y sustracción de ADN.

3. ALGORITMO

El cripto sistema desarrollado consiste de los algoritmos de cifrado y descifrado sobre imágenes digitales.

El algoritmo de cifrado consta de los pasos siguientes:

1. Se ingresa la imagen digital I de dimensión $m \times n$.
2. Se convierten los valores decimales de la imagen I a valores binario obteniendo I_b .
3. Se codifica la imagen binaria I_b a una secuencia de ADN y el resultado se deja en una matriz denominada L .
4. Se construye la llave de la forma siguiente:
Se calculan k_1 y k_2 usando las siguientes ecuaciones:

$$k_1 = \frac{1}{126} \bmod \left(\sum_{i=1}^{m/2} \sum_{j=1}^n a_{ij}, 256 \right)$$

$$k_2 = \frac{1}{126} \bmod \left(\sum_{i=m/2+1}^m \sum_{j=1}^n a_{ij}, 256 \right)$$

Ahora, para iniciar con la generación de la llave se obtienen los vectores: X, Y de la manera siguiente: Se generan de forma aleatoria dos valores: x_0, y_0 sobre el intervalo real $[0, 1]$. Se aplica el proceso siguiente para x_0 (se repite el mismo proceso pero usando y_0 en lugar de x_0) (Tere, Vaida, Alboai y Chiorea, 2011):

$$x_0 = x_1 + k_1$$

Si $x_0 > 1$ entonces

$$x_0 = \bmod(x_0, 1)$$

Si no

$$x_0 = x_0$$

Una vez obtenido los valores de x_0 y y_0 se procede a generar la llave $X = (x_1, \dots, x_{8m})$ como vector columna y $Y = (y_1, \dots, y_{8n})$ como vector fila con la siguiente formula caótica conocida como *cross chaotic map* (Tere, Vaida, Alboai y Chiorea, 2011):

$$x_{i+1} = 1 - ny_i^2$$

$$y_{i+1} = \cos(k \cos^{-1}(x_i)), \text{ donde } k \text{ es una constante.}$$

Se multiplica el vector columna X por el vector fila Y , obteniendo una matriz M de dimensión $8m \times 8n$.

5. Se convierte la matriz M a binario donde se pone el valor a 0 si $M(i, j) < 0$ y a valor 1 si $M(i, j) \geq 0$, el resultado se deja en la matriz M_b .
6. Se codifica M_b a cadenas de ADN obteniendo la matriz K .
7. Se aplica la adición de ADN como lo dicta el cuadro 1.
 $N_{ADN} = L + K$.
 La matriz obtenida N_{ADN} se pasa a binario obteniendo N_b .
8. Se convierte N_b a valores enteros agrupando cada 8 bits para formar un valor entero entre 0 y 255, obteniendo una matriz C de dimensión $m \times n$.

El resultado final es la matriz C , la cual es la imagen cifrada.

Para el algoritmo de descifrado se siguen los siguientes pasos:

1. Se pasa la matriz C a valores binarios, obteniendo C_b .
2. Se codifica C_b a valores de ADN obteniendo C_{ADN} .
3. Se aplica la regla de sustracción de ADN dada por el cuadro 2
 $L_2 = C_{ADN} - K$. Donde K es la llave generada en el proceso de cifrado.
4. Se convierte L_2 a binario obteniendo L_b .
5. Se convierte L_b a decimal, obteniendo I .

Al finalizar el algoritmo de descifrado se obtiene la imagen descifrada I que debe corresponder a la imagen original.

4. PRUEBAS

La implementación computacional se desarrolló bajo el lenguaje de programación OCTAVE [26] y las pruebas se realizaron en una computadora *hp* con 16 Gb de memoria RAM y un procesador intel core i7 a 2.6 GHz.

La figura 1a, muestra la imagen de un paciente que presenta la enfermedad de rubeola [4] [18], la imagen tiene una dimensión 210 x 280, es decir, 210 filas y 280 columnas. La rubeola es una enfermedad viral eruptiva, contagiosa y epidémica producida por un virus ultra filtrable. La figura 1b muestra el resultado del algoritmo de cifrado y la figura 1c muestra el resultado del algoritmo de descifrado.

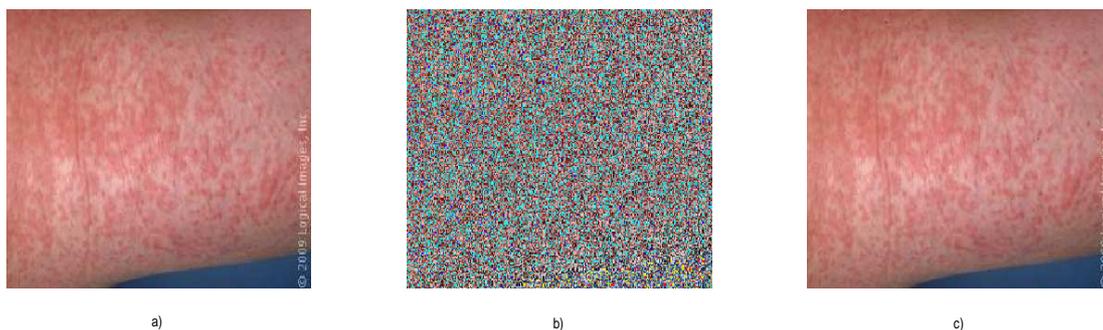


Figura 1. a) Imagen original de un paciente que presenta rubeola, b) Imagen cifrada, c) Imagen descifrada.

La figura 2a, muestra la imagen de un paciente que tiene la enfermedad de sífilis secundaria [25]. La imagen tiene dimensión de 600 x 503 píxeles. La sífilis es una enfermedad venérea contagiosa infectocontagiosa, provocada por la espiroqueta *Treponema pallidum*, que se manifiesta por un chancro y por

lesiones viscerales y encefálicas a largo plazo. La figura 2b muestra la imagen cifrada y la figura 2c muestra el descifrado.

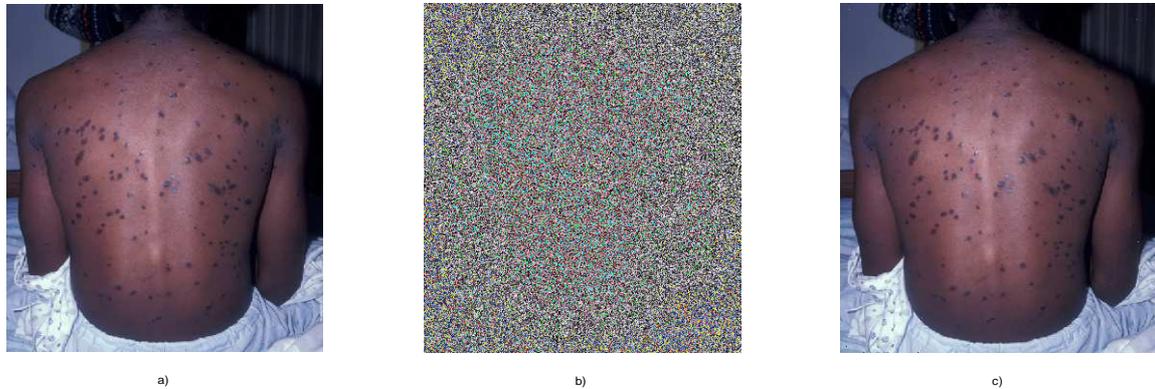


Figura 2. a) Imagen original de la enfermedad sífilis secundaria, b) Imagen cifrada, c) Imagen descifrada.

La figura 3a, muestra la imagen de un ultrasonido tiroideo en un paciente que presenta enfermedad en la tiroides, la imagen es de dimensión 327 x 430. Como su nombre lo indica el ultrasonido tiroideo es el estudio mediante ecografía de la Glándula Tiroides, como en todos los ultrasonidos no se emplea ningún tipo de radiación obteniéndose la imagen del eco de sonidos de alta frecuencia en el tejido estudiado [23]. La tiroides, es una glándula endocrina situada delante de la tráquea que produce la tiroxina y la calcitonina. La figura 3b muestra la imagen cifrada y la figura 3c muestra el descifrado.

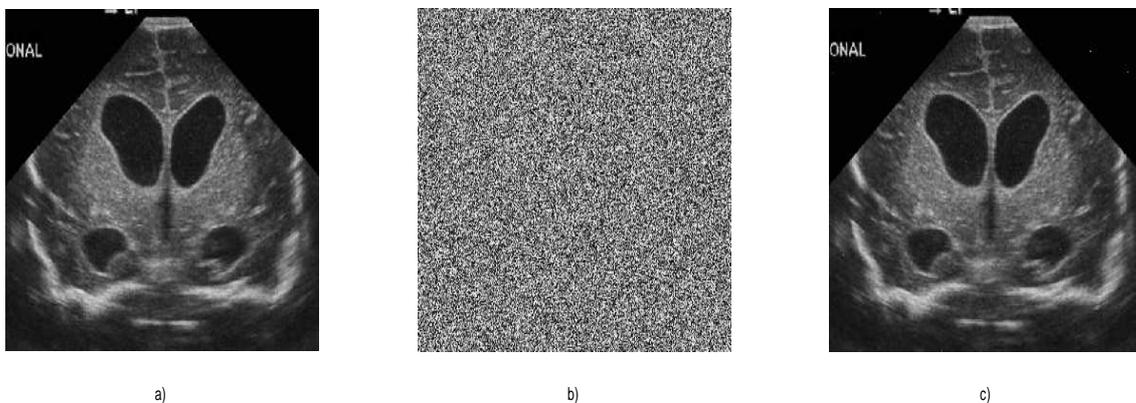


Figura 3. a) Imagen original de la enfermedad tiroides, b) Imagen cifrada, c) Imagen descifrada.

La figura 4a, muestra la imagen de un paciente que presenta una enfermedad en los ojos [24], la imagen es de dimensión 249 x 350. La figura 4b muestra la imagen cifrada y la figura 4c muestra el descifrado.

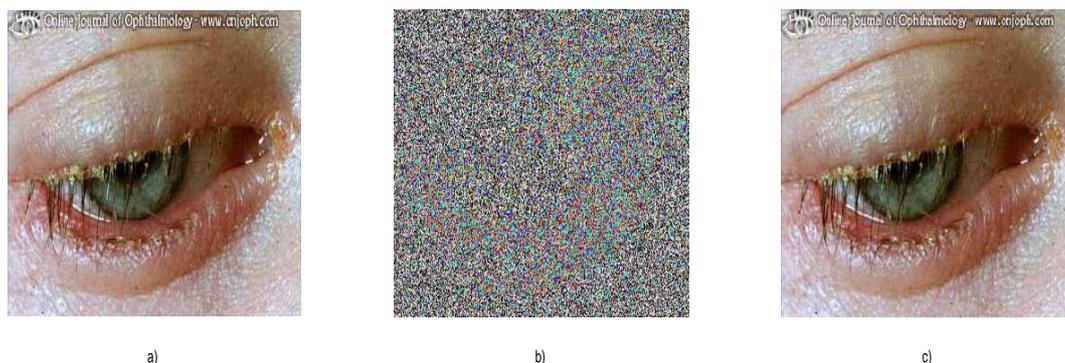


Figura 4. a) Imagen original de una enfermedad de los ojos, b) Imagen cifrada, c) Imagen descifrada.

5. CONCLUSIONES

El uso de las funciones caóticas, junto con las operaciones de adición y sustracción sobre ADN son útiles en el desarrollo de sistemas criptográficos simétricos, además de una alternativa rápida al estándar AES de la criptografía simétrica en grandes volúmenes de datos como lo requiere el caso del cifrado y descifrado de imágenes médicas.

Como se puede apreciar en las cuatro imágenes médicas utilizadas en este trabajo, los resultados de la implementación computacional del algoritmo de cifrado deja irreconocible el tipo de imagen que se trata y también se observa que el descifrado obtiene la imagen original.

REFERENCIAS

- [1] Anil, J. and Chirakkarott, S. (2014) Secure Encryption Method for Biometric Iris Pattern, International Journal of Trends and Technology (IJCTT), volume 12, number 5, 217-219.
- [2] Anwar, T., Sanchita, P. and Singh, S.K. (2014): Message Transmission Based on DNA Cryptography: Review, International Journal of Bio-Science and Bio-Technology, Vol.6, No.5, 215-222
- [3] Delfs, H., Maurer, U. and Knebl, H., (2007): Introduction to Cryptography Principles and Applications, Second Edition, Springer-Verlag Berlin Heidelberg.
- [4] Freedberg, I.M., (2003): Fitzpatrick's Dermatology In General Medicine. 6TH EDITION, New York: McGraw-Hill.
- [5] Fuan, Ch., Mengb, W., Zhanb, Y., Zhuc, Z., Laud, F., Tsed, Ch. and Mae, H. (2013) An efficient and secure medical image protection scheme based on chaotic maps. Computers in Biology and Medicine, Elsevier, vol. 43, 1000-1011.
- [6] Galende, J. (1995): Criptografía, historia de la escritura cifrada, 1ra edición, editorial complutense..
- [7] Gao, T.G. and Chen, Z.Q., (2008): Image encryption based on a new total shuffling algorithm, Chaos Solutions & Fractals, 38 (1), 213–220.
- [8] Goldreich, O. (1999): Modern Cryptography, theory and practice, discrete mathematics and its applications 3th edition, Chapman & Hall.
- [9] Hilborn, R.C. (2000): Chaos and nonlinear dynamics 2da edición, Oxford University Press.
- [10] Huang, C.K. and Nien, H.H., (2009): Multi chaotic systems based pixel shuffle for image encryption, Opt. Commun. , 282 (11), 2123–2127.
- [11] Imagen de una enfermedad en el ojo humano. Disponible en <http://www.atlasophthalmology.com/atlas/photo.jsf?node=3002&locale=de>, consultada 10-07, 2016.
- [12] Imagen Sífilis, Disponible en missinglink.ucsf.edu, consultada 10-07, 2016.
- [13] Javheri, S. and Kulkarni, R. (2014): Secure Data communication and Cryptography based on DNA based Message Encoding, International Journal of Computer Applications, Volume 98, No.16, 35-40.
- [14] Ott, E. (1993): Chaos in dynamical systems, Cambridge University Press.

- [15] Octave, Disponible en <https://www.gnu.org/software/octave/>, consultada 10-07, 2016.
- [16] Patidar, V., Pareek, N.K. and Sud, K.K. (2009): A new substitution–diffusion based image cipher using chaotic standard and logistic maps, *Commun. Nonlinear Sci. Numer. Simulation*, 14 (7) 3056–3075.
- [17] Rhouma, R., Meherzi, S. and Belghith, S. (2009): OCML-based colour image encryption, *Chaos Solitons & Fractals*, 40 (1), 309–318.
- [18] Sun, F.Y., Liu, S.T. and Li, Z.Q. (2008): A novel image encryption scheme based on spatial chaos map, *Chaos Solitons & Fractals*, 38 (3), 631–640.
- [19] Terece, R., Vaida, M.F., Alboaie, L. and Chiorea, L. (2011): DNA security using symmetric cryptography, *International journal of new computer architectures and their applications*, IJNCAA, vol. 1.
- [20] Tong, X.J. and Cui, M.G.,(2009): Image encryption scheme based on 3D baker with dynamical compound chaotic sequence cipher generator, *Signal Processing*, 89 (4) 480–491.
- [21] Ultrasonido Tiroideo. Disponible en <http://www.dei.org.mx/estudios/ultrasonido-4d-hd-live-y-doppler-color/ultrasonido-tiroideo/>, consultada 10-07, 2016.
- [22] Wollensak, J. (2003): Univ.-Augenlinik, Berlin, Germany.
- [23] Wong, K.W., Kwok, B.S. and Law, W.S. (2008): A fast image encryption scheme based on chaotic standard map, *Phys. Lett. A.*, 372 (15), 2645–2652.
- [24] Xiao Y.L. and Xia L.M. (2009): An Image Encryption Approach Using a Shuffling Map, *Commun. Theor. Phys.*, 52 (5) 876–880.
- [25] Xu, S.J., Wang, J.Z. and e S.X. (2008): An improved image encryption algorithm based on chaotic maps, *Chin. Phys. B*, 17 (11), 4027–4032.
- [26] Ye, G.D. (2009): A chaotic image cryptosystem based on Toeplitz and Hankel matrices, *Imaging Sci. J.*, 57 (5) 266–273.

PREDICCIÓN DE MORTALIDAD EN CUIDADOS INTENSIVOS ONCOLÓGICOS MEDIANTE REGRESIÓN LOGÍSTICA DE COMPONENTES PRINCIPALES CATEGÓRICOS

F. D. Martos-Benítez*, V. Sistachs Vega**, M. A. Díaz Martínez ***

* Unidad de Cuidados Intensivos 8B, Hospital Clínico Quirúrgico “Hermanos Ameijeiras”. La Habana, Cuba.

** Facultad de Matemática y Computación, Universidad de La Habana. La Habana, Cuba.

*** CEMAT “Centro de Estudios Matemáticos”, CUJAE “Ciudad Universitaria José Antonio Echeverría”. La Habana, Cuba.

ABSTRACT

Mortality in cancer patients admitted in intensive care unit (ICU) is high. Mortality predictive model for this type of patients is lack. Usually, these models are constructed using binary logistic regression (BLR) analysis; however, the relationship among predictive variables and which dimensions of predictive variables better predict the respond are not explained by this method. Statistical modelation using the categorical principal component analysis (CPCA), the BLR, and logistic regression based on the CPCA allows a better characterization of study population, identify the variables related with ICU mortality, and predict mortality with a high precision, as well as with a bigger efficiency and consistency.

KEYWORDS: critically ill cancer patient, oncological intensive care unit, mortality, binary logistic regression, categorical principal component analysis, categorical principal component analysis logistic regression.

RESUMEN

La mortalidad en los pacientes con cáncer admitidos en las unidades de cuidados intensivos (UCI) es alta. No existen modelos predictivos de muerte para este tipo de enfermos. Estos modelos usualmente se construyen mediante el análisis de regresión logística binaria (RLB); sin embargo, este método no explica las relaciones existentes entre las variables predictoras ni explica cuáles son las dimensiones de las variables predictoras que mejor predicen la respuesta. La modelación estadística mediante el empleo del análisis de componentes principales categóricos (ACPC), de la RLB y de la regresión logística basada en el ACPC, permite una mejor caracterización de la población en estudio, posibilita la identificación de las variables relacionadas con la mortalidad en UCI y predice la mortalidad con una alta precisión, así como con una mayor eficiencia y consistencia.

1. INTRODUCCIÓN

Clásicamente, los pacientes oncológicos han tenido menos oportunidades de acceder a las unidades de cuidados intensivos (UCI) por estigmatizaciones derivadas del pronóstico ominoso de la enfermedad. Aunque en los últimos años se ha apreciado un incremento en la aceptación de estos enfermos en esta área hospitalaria, la mortalidad continúa siendo elevada, reportándose cifras que superan el 50%. [1]

Los modelos predictivos de mortalidad en UCI, como el APACHE, el SAPS y el MPM, [2] se han desarrollado preferiblemente con sujetos no oncológicos o bien estos han representado una fracción muy pequeña de la población donde se realizaron, por lo que es dudosa la generalización y aplicación de estos modelos a los pacientes graves con cáncer. En tal sentido, se hace necesario construir modelos de predicción de la mortalidad específicamente para este tipo de pacientes, ya que constituyen una población con características particulares, no solo desde el punto de vista médico, sino también desde el punto de vista humano.

Antes de llevar a cabo la construcción de un modelo predictivo, es necesario caracterizar apropiadamente la población sujeto de estudio. Tradicionalmente se hace un análisis univariado, donde se estima si existen diferencias estadísticamente significativas entre el grupo de pacientes que falleció y el grupo de pacientes que sobrevivió durante el tiempo de seguimiento. Este análisis, si bien permite identificar posibles predictores, no es apropiado para definir correctamente las características de la población que se analiza, ni siquiera evalúa las relaciones existentes entre los predictores, que sería muy útil para un mejor entendimiento del problema a investigar. Una técnica apropiada para resolver este problema es el análisis de componentes principales (ACP).

La mayoría de las construcciones matemáticas de predicción se basan en modelos lineales o linealizados, donde se intenta predecir Y según los efectos aditivos de los predictores. Contextualizando, si se desea construir un modelo para predecir la probabilidad de morir en los pacientes con cáncer admitidos en UCI mediante un modelo de

regresión logística binaria (RLB), la probabilidad de morir ($Y = 1$) estaría dada por la sumatoria del múltiplo de un parámetro β_j por el valor ($X = x$) de las variables predictoras que se utilicen; sin embargo, este modelo no explica las relaciones que pueden existir entre las variables predictoras; tampoco explica cuáles son las dimensiones de las variables predictoras que mejor predicen $Y = 1$, lo cual permitiría ser más exhaustivo en la explicación del fenómeno que se estudia. Estas limitaciones también pueden ser resueltas mediante la modelación matemática empleando el ACP y el análisis de RLB.

Los objetivos del presente estudio fueron: 1-) caracterizar la población de pacientes con cáncer admitidos en UCIO oncológica; 2-) determinar los factores de riesgo de muerte en estos pacientes; y 3-) entre el modelo de RLB y el de regresión logística basado en el análisis de componentes principales categóricos (RLCPC), seleccionar el mejor modelo predictivo de muerte. Para lograr estos objetivos en el epígrafe 2 se plantea el diseño y las variables estudiadas y se muestra el procedimiento estadístico empleado; también se conceptualiza el ACP categóricos, el análisis de RLB, así como los criterios para evaluar la capacidad predictiva y los criterios para la selección de modelos. En el epígrafe 3 se presentan los resultados y la discusión de los mismos.

2. DISEÑO DEL ESTUDIO Y PROCEDIMIENTO ESTADÍSTICO

2.1 Diseño y contexto del estudio

Desde enero del 2014 hasta diciembre del 2015 se realizó un estudio de cohorte prospectivo en la unidad de cuidados intensivos oncológicos (UCIO) del Instituto de Oncología y Radiobiología (IOR) de Cuba. Durante el período que duró el estudio se incluyeron consecutivamente 1378 pacientes con cáncer. Se excluyeron los sujetos que tuvieron una estadía en UCIO ≤ 1 día, de manera que la muestra estuvo constituida por 961 individuos.

En las primeras 12 horas de ingreso en la UCIO se registraron las siguientes variables explicativas: edad, sexo, Índice de Comorbilidades de Charlson Ajustada para la Edad (con exclusión de la malignidad), localización del tumor primario, etapa clínica del cáncer, procedencia del paciente (sala hospitalaria o departamento de emergencias), admisión no relacionada con el tumor, tipo de admisión (médica o quirúrgica), escala *Acute Physiology and Chronic Health Evaluation* (APACHE) II, sepsis al ingreso en la UCIO y evento adverso a la quimioterapia como causa de ingreso en la UCIO.

La variable de respuesta que se analizó en el presente estudio fue la mortalidad en la UCIO.

2.2 Procedimientos estadísticos

2.2.1 Análisis univariado

En todas las variables explicativas se utilizaron medidas de resumen para describir los datos, de manera que las variables cualitativas se muestran en porcentaje; debido a la falta de normalidad, las variables cuantitativas se representan como mediana y 25% – 75% rango intercuantil (RIQ). La comparación entre el grupo de pacientes que falleció y el que sobrevivió en la UCIO, se realizó mediante la prueba de chi cuadrado (χ^2) o el test exacto de Fisher para las variables cualitativas, según fue apropiado; en las variables cuantitativas se empleó un proceder no paramétrico (prueba U de Mann-Whitney).

2.2.2 Análisis de regresión logística

Para identificar los factores de riesgo de muerte en UCIO se realizó un análisis de RLB multivariado. En la construcción del modelo de RLB se tiene una variable de respuesta o dependiente Y (en nuestro caso *mortalidad en UCIO*) que toma valores uno o cero, los cuales indican la presencia (*muerte*) o ausencia (*no muerte*) de ocurrencia del evento de interés, respectivamente. Siendo k el número de variables independientes del modelo, representadas por $\mathbf{X} = (X_1, X_2, \dots, X_k)^T$, la formulación genérica del modelo de RLB es $Y = P(\mathbf{x}) + \varepsilon$, donde ε es el término de error. Siendo $P(\mathbf{x})$ la probabilidad de que la respuesta Y tome el valor 1 para el valor observado \mathbf{x} ; el modelo se expresa como: [3]

$$P(Y = 1 | \mathbf{X} = \mathbf{x}) = P(\mathbf{x}) = \frac{\exp(\beta_0 + \sum_{i=1}^k \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^k \beta_i x_i)}$$

siendo $x = (x_1, x_2, \dots, x_k)^T$ el valor observado de las variables explicativas y β los coeficientes de regresión logística. Por tanto, $1 - P(x)$ indica la probabilidad de que Y tome el valor 0. Si se aplica una transformación logit a la ecuación anterior, se obtiene un modelo de regresión lineal que facilita la posterior interpretación del modelo:

$$\text{logit}(P(x)) = \ln \left(\frac{P(x)}{1 - P(x)} \right) = \beta_0 + \sum_{i=1}^k \beta_i x_i$$

En el modelo inicial se incluyeron las variables que tuvieron un p -valor bilateral $\leq 0,05$ en el análisis univariado. Debido a la falta de normalidad, las variables cuantitativas fueron categorizadas para este propósito. Se aplicó el método de selección automática de variables “paso a paso” para simplificar la complejidad del modelo final. Para este procedimiento se utilizó un p -valor para entrar de 0,05 y un p -valor para salir de 0,10. En los resultados se muestran los coeficientes de regresión para cada variable, el error estándar, el estadístico de Wald, el odds ratio (OR) con intervalo de confianza (IC) al 95% y el p -valor.

Un elemento central en el proceso de crear un modelo de RLB es su evaluación en cuanto al rendimiento. En este sentido, se han propuesto diversas medidas, que pueden agruparse en dos categorías principales: medidas de calibración y de discriminación. Independientemente del objetivo para el que se ha creado el modelo, estas dos medidas del rendimiento deben derivarse de los datos que le han dado origen.

La *calibración* es una medida que expresa la concordancia entre los resultados observados y las predicciones del modelo. En otras palabras, es la capacidad del modelo de producir estimaciones no sesgadas de la probabilidad del evento. La medida de calibración más habitual es la prueba de Hosmer-Lemeshow. [4]

Cuando los patrones de covariables siguen una distribución n -asintótica, la χ^2 residual de Pearson y la devianza de residuos no se ajustan bien a una distribución χ^2 . Por ello, en general, será más apropiada la prueba de Hosmer-Lemeshow, que agrupa a los n sujetos en m patrones según criterios estadísticos. En concreto, los sujetos se agrupan según 9 deciles de las probabilidades esperadas; a partir de aquí, se crea una tabla de contingencia de 10×2 de la que se construye un estadístico que seguirá una distribución χ^2 con 8 grados de libertad. La ausencia de significación indica un buen ajuste del modelo. [5]

Es conveniente comprobar la tabla de contingencia de la que deriva el estadístico. La presencia de celdas con frecuencias esperadas menores de 5 aconseja colapsar filas para eliminar estas celdas de baja frecuencia esperada; esto implicaría además reducir proporcionalmente los grados de libertad de la χ^2 empleada (1 por fila eliminada) y recalcular el estadístico.

La *discriminación* es la capacidad del modelo de asignar el resultado correcto a un par de sujetos seleccionados al azar; en otras palabras, permite al modelo clasificar a los sujetos en un contexto de criterio de valoración con predicción binaria. [4] El área bajo la curva de las características operativas del receptor (COR) es la medida de discriminación que se utiliza con mayor frecuencia en modelos de error normal y resultado binario.

Una curva COR enfrenta en un sistema de ejes la sensibilidad en el eje y , complementado con la especificidad en el eje x . El procedimiento consiste en determinar las correspondientes tablas de clasificación de puntos de corte de $P(Y=1|X)$ crecientes (0,1, 0,2, ..., 0,8, 0,9, 1) y determinar a partir de ellas las correspondientes sensibilidades y especificidades. Si el interés es el punto de corte óptimo para predecir Y , es decir, el punto de corte que ofrece mayor sensibilidad y especificidad, se recurre al análisis gráfico de la curva, seleccionando como punto de corte aquél que corresponde con el punto de inflexión de la curva COR. [6]

Diferentes modelos ofrecen diferentes curvas COR. La comparación entre modelos respecto a su capacidad de discriminación puede hacerse comparando el área bajo las curvas COR; las mejores curvas serán aquellas con un área más próxima a la unidad. Como regla general, un área de 0,5 implica ausencia de discriminación; entre 0,7 y 0,79 es una discriminación aceptable; entre 0,8 y 0,89 es excelente; 0,9 o superior es una discriminación excepcional.

2.2.3 Análisis de componentes principales categóricos

Se realizó un análisis de componentes principales categóricos (ACPC) mediante el método de escalamiento óptimo. Este análisis es una generalización del ACP lineal, que busca maximizar la varianza total de las r primeras componentes principales, es decir, la traza de la matriz de covarianzas de las p componentes principales.

El método de ACPC puede considerarse como una técnica exploratoria de reducción de las dimensiones incorporando variables nominales y ordinales de la misma manera que las numéricas. [7] El método pone al descubierto relaciones existentes entre las variables originales, entre los casos y entre ambos (variables y casos). Puede, además, analizar variables con su nivel de medición. Cuando existe relación no lineal entre las variables, pueden especificarse también otros niveles de análisis, de manera que estas relaciones pueden manipularse de forma más efectiva. El modelo de ACPC es igual al modelo del ACP lineal, capturando las posibles no linealidades de las

relaciones entre las variables originales en las transformaciones de las mismas. El objetivo del ACP se alcanza por el ACPC minimizando la función de pérdida; en esta función se acomodan las ponderaciones de acuerdo con los valores ausentes, ponderaciones por casos y transformaciones nominales múltiples. [8]

De la matriz de variables categóricas X , es necesario transformar en numéricas las categorías para realizar el ACP, para lo cual se recurre al proceso iterativo de mínimos cuadrados alternantes (ALS, por sus iniciales en inglés). Cada iteración del ALS tiene dos etapas: 1-) cuantificación óptima en una matriz de cuantificaciones denotada por Q para maximizar las correlaciones entre las variables y que contiene las variables transformadas $q_j = \phi_j(y_j)$; [22] y 2-) la estimación del modelo de ACP lineal sobre la matriz Q cuantificada. [9]

Con la valoración de las variables cualitativas por el procedimiento de ALS, se logra también maximizar la correlación lineal entre las variables originales, consiguiendo con esto transformar las variables cualitativas en variables cuantitativas, disponiendo así de la mejor combinación lineal posible de las variables originales.

El objetivo que se persiguió en este estudio con el ACPC no fue reducir dimensiones, sino emplearlo para construir un modelo predictivo de regresión logística. En tal sentido, la pérdida de información propia de la reducción de dimensiones podría afectar la predicción, por lo que se utilizaron todas las dimensiones en el análisis, en total nueve.

Se evaluó el alfa de Cronbach, los valores propios y el porcentaje de la varianza explicado por cada dimensión. Para caracterizar la población sujeto de estudio se utilizaron las dimensiones que tuvieron valores propios mayores de uno y un alfa de Cronbach positivo. Se examinaron las cargas de las variables en cada dimensión (incluyendo el análisis gráfico) y se consideraron significativas las variables con cargas mayores o iguales a 0,40, independientemente del signo. Se construyó el gráfico conjunto de las categorías para las dimensiones empleadas para caracterizar a la población y para establecer posibles relaciones entre las variables y las categorías de las variables.

Las puntuaciones de los objetos (pacientes) en los componentes (dimensiones) fueron salvadas en la base de datos para usarlas como variables explicativas (*variables ficticias*) en un modelo de RLCPC. En este modelo se procedió de la misma forma que en el modelo de RLB.

2.2.4 Comparación de modelos de regresión logística

Ambos modelos, el de RLCPC y el de RLB, se compararon en base a los siguientes criterios: prueba de $H - L$, área bajo la curva COR, Criterio de Información de Akaike (AIC) y Criterio de Información Bayesiano (BIC). Con la prueba de $H - L$ solo se buscó el ajuste de los modelos. Se consideró como mejor modelo aquel que tenía una mayor área bajo la curva COR y menores valores en el AIC y el BIC.

Todo el análisis estadístico se realizó mediante el programa para análisis estadístico IBM® SPSS® Statistics 20.0 (IBM, Chicago, IL, USA).

3. RESULTADOS Y DISCUSIÓN

3.1 Características de los sujetos, análisis univariado y análisis de componentes principales categóricos

Se estudiaron 691 pacientes con cáncer ingresados en la unidad de cuidados intensivos oncológicos (UCIO). La distribución de sexo fue similar (*femenino* 51,2% vs. *masculino* 48,8%) y la edad tuvo una mediana de 61,0 años (RIQ 52,0 – 69,0 años). La mayoría de los enfermos tenía morbilidades crónicas, obteniéndose dos o más puntos en la escala de Charlson Ajustada para la Edad en el 86,4%. Los tumores primarios se localizaron con mayor frecuencia en el tracto gastrointestinal (35,0%) o en la región del tórax (24,7%). Predominó el ingreso por causas quirúrgicas (78,6%). En el 81,8% de los pacientes la admisión en UCIO se debió a un trastorno relacionado con el proceso neoplásico. El 90,3% de los sujetos procedía de las salas hospitalarias y la sepsis estaba presente en el 7,8%. Los eventos adversos a las drogas empleadas para la quimioterapia representaron el 5,1% de los motivos de ingreso en UCIO. La mediana de puntuación en la escala APACHE II al ingreso en la UCIO fue de 12,0 puntos (RIQ 10,0 – 14,0 puntos). El 15,8% de los pacientes requirió ventilación artificial mecánica (VAM) durante su estancia en la UCIO. En la UCIO fallecieron 97 individuos, lo que representó una incidencia del 14,04%.

En la tabla 1 se muestran las características entre los pacientes que fallecieron y los que se egresaron vivos. En el análisis univariado se observaron diferencias estadísticamente significativas entre los grupos en relación a la escala de Charlson Ajustada para la Edad, la localización del tumor primario, el tipo de paciente, el ingreso en UCIO no relacionado con el proceso tumoral, la procedencia de los pacientes, la presencia de sepsis al ingreso en la UCIO,

los eventos adversos a los quimioterápico como causa de ingreso en la UCIO, la puntuación en la escala APACHE II al ingreso en la UCIO y la necesidad de VAM.

Los resultados del análisis univariado indican las posibles variables que pueden constituir factores de riesgo de muerte en UCI y que serán incluidas en el análisis multivariado de RLB. Sin embargo, este análisis no es apropiado para una adecuada caracterización de la población sujeto de estudio, ni permite identificar las relaciones existentes entre los predictores y las categorías de las variables. Para ello se realizó un ACPC.

Tabla 1. Características de los pacientes que fallecieron y los que fueron egresados vivos.

Variables	Muertos N= 97	Vivos N= 594	P
Edad, años [mediana (RIQ)]	63,0 (52,0-72)	61,0 (51,0-69,0)	0,205
Edad, n (%)			0,554
< 30 años	3 (3,1)	23 (3,9)	
30-39 años	4 (4,1)	29 (4,9)	
40-49 años	9 (9,3)	73 (12,3)	
50-59 años	27 (27,8)	151 (25,4)	
60-69 años	24 (24,7)	180 (30,3)	
≥ 70 años	30 (30,9)	138 (23,2)	
Sexo, n (%)			0,879
Masculino	48 (49,5)	289 (48,7)	
Femenino	49 (50,5)	30,5 (51,3)	
Charlson Ajustado, puntos [mediana (RIQ)]	4,0 (3,0-6,0)	3,0 (2,0-4,0)	<0,0001
Charlson Ajustado, n (%)			0,001
≤ 1 punto			
2-3 puntos	6 (6,2)	88 (14,8)	
4-5 puntos	30 (30,9)	246 (41,4)	
≥ 6 puntos	34 (35,1)	176 (29,6)	
Tumor Primario, n (%)			<0,0001
Cabeza y Cuello	7 (7,2)	50 (8,4)	
Tórax	17 (17,5)	154 (25,9)	
Gastrointestinal	21 (21,6)	221 (37,2)	
Ginecológico	16 (16,5)	47 (7,9)	
Urológico	9 (9,3)	30 (5,1)	
Hematológico	13 (13,4)	27 (4,5)	
Cerebral	3 (3,1)	13 (2,2)	
Mama	6 (6,2)	16 (2,7)	
Otros	5 (5,2)	36 (6,1)	
Tipo de Paciente, n (%)			<0,0001
Médicos	60 (61,9)	88 (14,8)	
Quirúrgicos	37 (38,1)	506 (85,2)	
Ingreso no Relacionado con el Tumor, n (%)	51 (52,6)	75 (12,6)	<0,0001
Procedencia, n (%)			<0,0001
Unidad de Urgencias	26 (26,8)	41 (6,9)	
Sala Hospitalaria	71 (73,2)	553 (93,1)	
Sepsis, n (%)	23 (23,7)	31 (5,2)	<0,0001
Evento Adverso QTP, n (%)	17 (17,5)	18 (3,0)	<0,0001
APACHE II, puntos [mediana (RIQ)]	12,7 (10,5-18,0)	12,0 (10,0-13,7)	<0,0001
APACHE II, n (%)			<0,0001
< 10 puntos	13 (13,4)	128 (21,5)	
10-14 puntos	41 (42,3)	369 (62,1)	
15-19 puntos	25 (25,8)	89 (15,0)	
≥ 20 puntos	18 (18,6)	8 (1,3)	
VAM, n (%)	59 (60,8)	50 (8,4)	<0,0001

APACHE, *Acute Physiology and Chronic Health Evaluation*; QTP, quimioterapia; RIQ, rango intercuartil; VAM, ventilación artificial mecánica.

Fuente: Datos de la investigación.

En la tabla 2 se representa el resumen del modelo del ACPC, que incluye los valores propios para cada dimensión y el alfa de Cronbach asociado. Dado que todas las variables fueron escaladas como ordinales o nominales, la varianza total explicada equivale a la sumatoria de los valores propios de las dimensiones. Los valores propios muestran el porcentaje de información retenida para cada dimensión, que en el ACPC cuando las variables se

escalas como ordinales o nominales, es equivalente a los valores propios del análisis de componentes principales lineal. [10]

Tabla 2. Resumen del análisis de componentes principales categóricos.

Dimensiones	Alfa de Cronbach	Varianza Explicada	
		Total (valores propios)	% de la Varianza ^b
1 ^c	,754	3,033	33,698
2 ^c	,104	1,102	12,243
3	-,020	,982	10,913
4	-,075	,938	10,420
5	-,179	,863	9,586
6	-,325	,776	8,622
7	-,705	,615	6,831
8	-1,249	,474	5,266
9	-4,036	,218	2,422
Total	1,000 ^a	9,000	100,000
a. El alfa de Cronbach total se basa en los valores propios totales			
b. En negrita se resaltan los componentes donde se alcanzó un porcentaje acumulado de la varianza explicada > 75%			
c. Componentes donde se obtuvieron valores propios > 1 y un alfa de Cronbach positivo			

Fuente: Datos de la investigación.

Se aprecia que los primeros cinco componentes acumularon el 76,86% de la información; sin embargo, sólo los dos primeros componentes tuvieron valores propios mayores a uno y un alfa de Cronbach positivo. Persiguiendo como objetivo caracterizar la población en estudio, estos dos primeros componentes fueron los elegidos, siempre teniendo en consideración que en ellos solo se explica el 45,9% de la variabilidad de los datos.

Varias técnicas se han propuesto para la selección del número de componentes a extraer [11,12]. Uno de los métodos más utilizado es el de Kaiser-Guttman, donde se extraen los componentes con valores propios mayores a la unidad, ya que en los mismos se concentra la mayor cantidad de la información. Este suele ser el método por defecto empleado en muchos de los paquetes estadístico de uso general; sin embargo, el desempeño puede ser pobre [11] debido a que datos incorrelacionados podrían aleatoriamente dar valores propios mayores a uno [12].

Otra técnica popular es el método de Cattell, consistente en el examen visual del gráfico de valores propios y en consecuencia, es relativamente subjetivo y requiere experiencia por parte del investigador [11]. Otro método propuesto para determinar el número de componentes es el porcentaje acumulado de la varianza total explicada. En este método, el número de componentes principales necesarios para describir adecuadamente los datos se basa en la porción de la varianza explicada, usualmente un 50-75% de la varianza acumulada [12]. Tiene el inconveniente que si se extraen más de tres componentes entonces es difícil el análisis gráfico de los mismos.

En algunos estudios de bioinformática se ha sugerido que el primer o los dos primeros componentes son suficientes para explicar los datos [13]; sin embargo, este enfoque se basa en observaciones empíricas del análisis de un número pequeño de bases de datos y carece de justificaciones estadísticas sólidas.

De la revisión de la literatura se infiere, que la elección del número adecuado de componentes es un aspecto aún en controversia. Nosotros apoyamos la opinión de Ma y Kosorok que dice [14]: “en el análisis práctico de datos, nosotros sugerimos que los investigadores exploren diferentes números de componentes principales y seleccionen el número adecuado basado en, por ejemplo, las implicaciones biológicas y el poder predictivo ... (en el análisis de regresión)”.

Nosotros, de preferencia, utilizamos el método de Kaiser-Guttman por ser uno de los más empleados y por lo tanto, mejor definidos, además de que tiene un mejor sustento estadístico en sus basamentos.

En la figura 1 se muestran las cargas de las variables para los dos primeros componentes. Las variables que mayor correlación tuvieron dentro del componente 1 fueron: tipo de paciente, ingreso relacionado con causa no tumoral, procedencia, infección al ingreso en UCIO y evento adverso a la quimioterapia como causa de ingreso en la UCIO. Dentro del componente 2, la escala de Charlson Ajustada para la Edad y la escala APACHE II fueron las más correlacionadas. Por tanto, podemos bautizar al primer componente como “características generales” y al segundo componentes como “escalas de riesgo”.

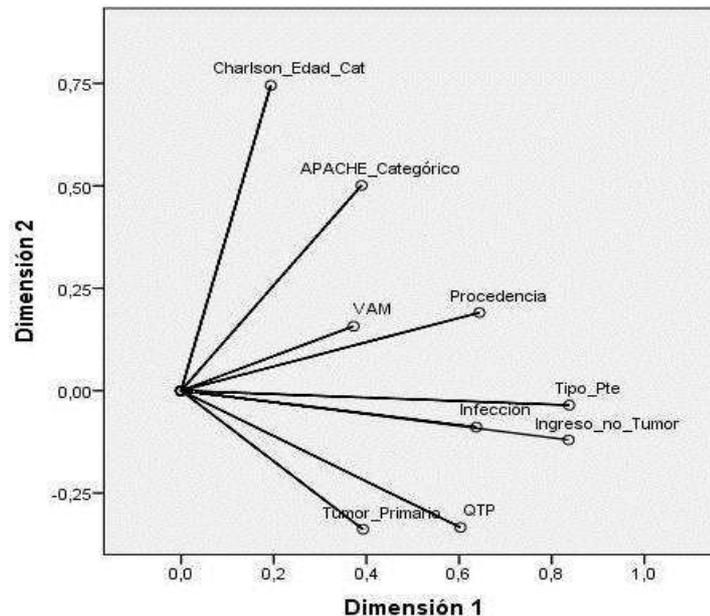


Figura 1. Carga de las variables en los componentes en el análisis de componentes principales categóricos.

Fuente: Datos de la investigación.

Considerando la proximidad de los vectores de las variables, en la figura 1 también se puede apreciar que existe una fuerte correlación entre el tipo de paciente, el ingreso en UCIO por causa no tumoral y la presencia de infección al ingreso en la UCIO, así como entre la VAM y la procedencia; también la escala APACHE II y la escala de Charlson Ajustada para la Edad se correlacionan, así como el evento adverso a la quimioterapia como causa de ingreso en la UCIO y la localización del tumor primario.

En la figura 2 se muestra el gráfico conjunto de los puntos de las categorías para los dos primeros componentes, donde se aprecia que las categorías que tuvieron una alta correlación en la dimensión 1 fueron los eventos adversos a la quimioterapia como causa de ingreso en la UCIO, la presencia de infección al ingreso en la UCIO, la procedencia del departamento de emergencia, el ingreso no relacionado con el tumor y los pacientes ingresados por causa médica. En la dimensión 2, las categorías que más se correlacionaron fueron la puntuación en la escala APACHE II ≥ 20 y la puntuación en la escala de Charlson Ajustada para la Edad ≥ 6 .

La exploración de las relaciones existentes entre las variables predictoras no es posible en un análisis de regresión logística, lo cual es importante porque permite una mejor comprensión del fenómeno que se investiga y una explicación más exhaustiva del problema. Ello facilita, además, considerar nuevas hipótesis para futuras exploraciones. Por otra parte, este análisis podría ayudar en la selección de las mejores variables predictoras para el análisis de RLB; por ejemplo, dada la fuerte correlación entre el tipo de paciente, la infección al ingreso en la UCIO y el ingreso en UCIO por causa no relacionada con el tumor (ver figura 1), en lugar de incluir todas estas variables en un modelo de RLB, podríamos considerar solo la inclusión de las que más se asociaron con el evento de interés en el análisis univariado o las de mayor importancia clínica. Así, podría quitarse la variable “infección al ingreso en UCIO” ya que probablemente la mayoría de los pacientes que ingresan por causa no relacionada con el tumor lo hace por infección. También podría considerarse no seleccionar la variable “VAM” o la variable “procedencia”. Como desde el punto de vista médico la VAM es una condición de mayor peso, podríamos elegir esta sobre la otra. Cuando analicemos el modelo de RLB, veremos que mediante la selección automática de variables, tanto la variable “infección al ingreso en UCIO” como la “procedencia”, fueron eliminadas y no permanecieron en el modelo final.

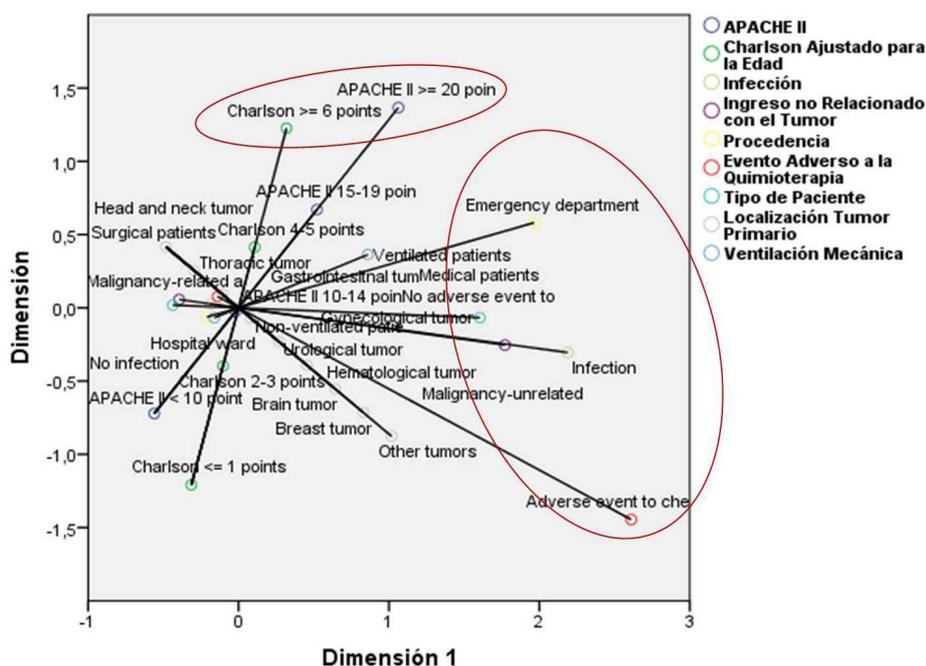


Figura 2. Gráfico conjunto de los puntos de las categorías para los dos primeros componentes en el análisis de componentes principales categóricos.

Fuente: Datos de la investigación.

3.2 Regresión logística binaria

Para el análisis de RLB se incluyeron las variables originales que resultaron significativas en el análisis univariado. En el modelo final (tabla 3), los pacientes admitidos en UCIO por causa médica, el uso de VAM, una escala APACHE II ≥ 20 puntos y una escala de Charlson Ajustada para la Edad ≥ 2 puntos son factores de riesgo independientes de muerte en la UCIO. La variable que tuvo un mayor efecto sobre la mortalidad fue la VAM, con un OR de 17,52 (95% IC 9,50 – 32,33; $p < 0,0001$).

Tabla 3. Modelo final de regresión logística binaria.*

Variables	β	ES	Wald	gl	P	OR	95% IC para OR	
							Inferior	Superior
Tipo de Paciente	1,442	,437	10,879	1	,001	4,228	1,795	9,957
VAM	2,864	,312	84,002	1	,000	17,523	9,499	32,325
APACHE II < 10 pts			10,265	3	,016			
APACHE II 10-14 pts	-,190	,397	,230	1	,631	,827	,380	1,799
APACHE II 15-19 pts	,454	,450	1,020	1	,313	1,575	,652	3,805
APACHE II ≥ 20 pts	1,511	,641	5,550	1	,018	4,531	1,289	15,928
Charlson-Edad ≤ 1 pto			11,209	3	,011			
Charlson-Edad 2-3 pts	1,310	,565	5,381	1	,020	3,708	1,225	11,220
Charlson-Edad 4-5 pts	1,763	,571	9,540	1	,002	5,828	1,904	17,837
Charlson-Edad ≥ 6 pts	1,815	,595	9,318	1	,002	6,140	1,915	19,690
Ingreso no Tumoral	,856	,442	3,759	1	,053	2,354	,991	5,594
Constante	-4,977	,654	57,982	1	,000	,007		

APACHE, escala *Acute Physiology and Chronic Health Evaluation*; ES, error estándar; gl, grados de libertad; IC, intervalo de confianza; OR, odds ratio; VAM, ventilación artificial mecánica.

* Variables incluidas en el modelo inicial: escala APACHE II, escala de Charlson Ajustada para la Edad, eventos adversos a la quimioterapia como causa de ingreso en UCIO, ingreso en UCIO no relacionado con el tumor, localización del tumor primario, presencia de sepsis al ingreso en la UCIO, procedencia del paciente, tipo de paciente y VAM.

Fuente: Datos de la investigación.

El modelo final de RLB tuvo una adecuada bondad de ajuste (prueba de H – L; $X^2= 6,550$; $p= 0,477$) y la discriminación fue muy buena, con un área bajo la curva COR de 0,912 (95% IC 0,882 – 0,941; $p < 0,0001$).

La identificación de estos factores de riesgo de muerte en UCIO permite una mejor estratificación de riesgo y evaluar el pronóstico de los pacientes con mayor objetividad. La mayoría de estos factores no son modificables, por lo que los sujetos en riesgo deben recibir una atención médica diferenciada y una monitorización intensiva estricta para prevenir, en la medida de lo posible, la probabilidad de morir.

3.3 Análisis de regresión logística de componentes principales categóricos

Tanto en la RLB como en la RLCPC, el objetivo de la selección de variables es elegir las mejores variables que expliquen $Y = 1$, lo que permite simplificar la complejidad del modelo estadístico. Sin embargo, en el modelo final de RLB se encuentran $r < p$ variables, donde p es el número de variables incluidas en el modelo inicial. Por su parte, en el análisis de RLCPC se trabaja con *variables explicativas ficticias* que son combinaciones lineales de las variables originales; así, aunque en el modelo final se encuentren $r < p$ *variables explicativas ficticias*, en r están representadas todas las p variables originales. Por tanto, en realidad, mediante la selección de variables en la RLCPC se eligen las “mejores dimensiones” de las variables originales que explican $Y = 1$.

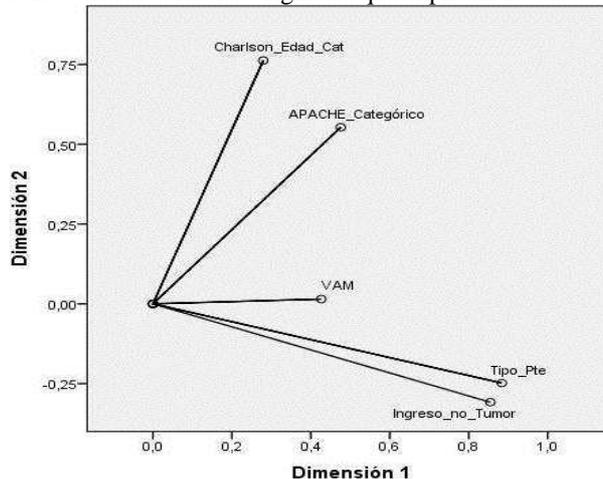


Figura 3. Carga de las variables en los componentes en el análisis de componentes principales categóricos del modelo reducido.

Fuente: Datos de la investigación.

Si realizamos un análisis de RLCPC considerando el ACPC previo (donde se incluyeron nueve variables y en consecuencia nueve combinaciones lineales, o sea, nueve componentes principales), en el modelo inicial y en el modelo final estarían representadas las nueve variables originales, mientras que en el modelo final de RLB solo se representan cinco de las variables originales. Por ello, el modelo final de RLCPC solo sería comparable con el modelo inicial de RLB, ya que en estos dos modelos es donde se encuentran representadas las p variables originales. Como en los resultados de la RLB solo son de interés las variables que quedaron en el modelo final, para evitar esta problemática se realizó un nuevo ACPC donde solo se incluyeron las cinco variables que quedaron en el modelo final de RLB. Ma y Kosorok apreciaron que cuando se utiliza el ACP para construir un modelo predictivo, la reducción de la dimensionalidad puede afectar la capacidad predictiva; [14] por tanto, el objetivo de este análisis no fue la reducción de las dimensiones, tomándose como *variables explicativas ficticias* para la RLCPC las cinco componentes extraídas.

Los tres primeros componentes explicaron el 80% de la varianza, aunque solo los dos primeros tuvieron autovalores mayores a 1 y un alfa de Cronbach positivo. En la figura 3 se representa la carga de las variables en los dos primeros componentes. En esta misma figura se puede apreciar la fuerte correlación que existe entre la variable “tipo de paciente” y la variable “ingreso por causa no relacionada con el tumor”. Esta última variable no resultó significativa en el modelo final de RLB (ver tabla 3), lo que apoya la conveniencia de utilizar el ACPC como un método para la evaluación de qué variables a incluir en un análisis de RLB.

En la tabla 4 se muestra el modelo final del análisis de RLCPC; los cuatro primeros componentes se asociaron significativamente con la mortalidad en UCIO, siendo poderosa la relación del primer componente (ingreso médico no tumoral de alto riesgo).

Tabla 4. Modelo reducido de regresión logística de componentes principales categóricos.

Variables	β	ES	Wald	gl	P	OR	95% IC para OR	
							Inferior	Superior
Componente 1	1,511	,145	108,191	1	,000	4,531	3,409	6,024
Componente 2	,300	,129	5,425	1	,020	1,350	1,049	1,737
Componente 3	,539	,102	28,098	1	,000	1,714	1,404	2,091
Componente 4	,462	,124	14,000	1	,000	1,588	1,246	2,023
Constante	-2,669	,189	199,606	1	,000	,069		

ES, error estándar; gl, grados de libertad; IC, intervalo de confianza; OR, odds ratio.

Fuente: Datos de la investigación.

Según la prueba de H – L, el modelo final de RLCPC tuvo una buena bondad de ajuste (X^2 7,907; $p=$ 0,341). Igualmente, la capacidad de discriminación de este modelo fue muy buena, con un área bajo la curva COR de 0,907 (95% IC 0,877 – 0,936; $p <$ 0,0001).

3.4 Comparación de modelos

Al comparar la prueba de H – L del modelo final de RLBCPC con la del modelo final de RLB, se observó que el modelo final de RLB ajustó mejor que el modelo final de RLCPC. Sin embargo, hay que tener en cuenta que la prueba de H – L no es apropiada para la comparación de modelos de regresión logística, lo importante es que se garantice una adecuada bondad de ajuste de los modelos, [15] lo que se observó en ambos modelos. También se debe considerar que la prueba de H – L casi siempre arroja una buena bondad de ajuste con tamaños muestrales grandes [16], argumento que contribuye a no centrar la comparación de ambos modelos basado en esta prueba.

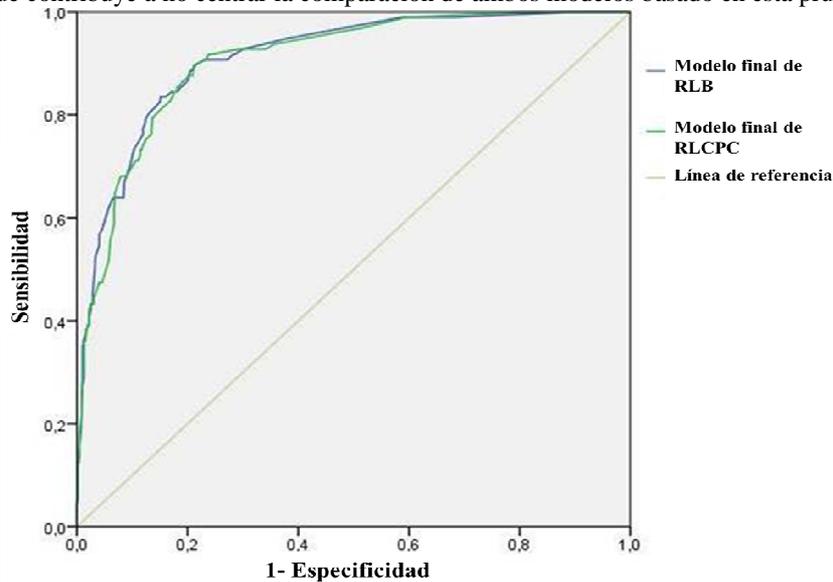


Figura 3. Curva de las características operativas del receptor del modelo final de regresión logística binaria y del modelo final de regresión logística de componentes principales categóricos.

RLB, regresión logística binaria típica; RLCPC, regresión logística binaria de componentes principales categóricos.

Fuente: Datos de la investigación.

Las características de las curvas COR y de las áreas bajo las mismas fueron casi idénticas entre el modelo de RLB y el modelo de RLCPC (figura 3). La curva COR provee una información estadística y visual que ha facilitado su uso corriente en los estudios biomédicos [17]. Es un método muy útil para comparar la capacidad de discriminación de los modelos predictivos de regresión logística [18].

Austin y Steyerberg observaron, que en las variables explicativas continuas la capacidad de discriminación del estadístico c , equivalente al área bajo la curva COR, no debe juzgarse solo por el odds ratio, ya que la heterogeneidad de la varianza entre el grupo de sujetos con el evento de interés y el grupo sin el evento puede afectar los resultados [19]. Esto debe tenerse en cuenta, porque nosotros comparamos un modelo estrictamente basado en variables cualitativas (modelo de RLB) con un modelo basado estrictamente en variables continuas (modelo de RLCPC); en tal sentido, probablemente la capacidad de discriminación del modelo de RLCPC sea infravalorada por la evaluación del área bajo la curva COR.

Al comparar ambos modelos tomando en cuenta las pruebas para la selección de modelos, como se aprecia en la tabla 5, el AIC y el BIC fueron menores en el modelo final de RLCPC que en el modelo final de RLB.

Tabla 5. Comparación de las pruebas para la selección de modelos entre el modelo final de regresión logística de componentes principales categóricos y el modelo final de regresión logística binaria.

Modelos	Criterio de Información de Akaike	Criterio de Información Bayesiano
Modelo final de RLCPC	169,13	191,82
Modelo final de RLB	171,95	217,33

RLB, regresión logística binaria; RLCPC, regresión logística de componentes principales categóricos.

Fuente: Datos de la investigación.

En virtud de la buena bondad de ajuste y de la casi idéntica capacidad de discriminación de ambos modelos, pero con un menor valor en el BIC y en el AIC en el modelo final de RLCPC respecto al modelo final de RLB, podemos afirmar que es más conveniente seleccionar el primero.

También es importante destacar, que para una adecuada interpretación de los resultados ofrecidos por el análisis de RLCPC, es necesario la ejecución de un ACPC y la identificación de las dimensiones que caracterizan a la población estudiada. Tales características permitirán identificar los sujetos en riesgo y poder tomar las medidas médicas apropiadas para prevenir un desenlace fatal en los pacientes. Cada componente del ACPC puede interpretarse como un índice formado para las combinaciones lineales de las variables.

Cuando se construya un modelo predictivo con variables categóricas, consideramos prudente el modelado estadístico con el empleo de la RLB y de la RLCPC, ya que la primera posibilita identificar las variables que se relacionan de forma independiente con la variable de respuesta. Por su parte, la RLCPC permite identificar qué índice (o combinación lineal) se relaciona con la respuesta, además que se demostró que la RLCPC es un modelo que ajusta mejor que la RLB y con una similar capacidad de discriminación.

4. CONCLUSIONES Y RECOMENDACIONES

Dados los resultados obtenidos, se puede concluir que: 1) el ACPC permite identificar las dimensiones que caracterizan a la población, tales como: *Dimensión 1 (características generales)*: tipo de paciente, ingreso relacionado con causa no tumoral, procedencia, infección al ingreso en UCIO y evento adverso a la quimioterapia como causa de ingreso en la UCIO; y *Dimensión 2 (escalas de riesgo)*: Escala APACHE II ≥ 15 puntos y escala de Charlson Ajustada para la Edad ≥ 6 puntos; 2-) el ingreso en UCI por causa médica, la necesidad de VAM, la puntuación en la escala APACHE II ≥ 20 y la puntuación en la escala de Charlson Ajustada para la Edad ≥ 2 son factores de riesgo independientes de muerte en UCI; 3-) aunque la capacidad predictiva de ambos modelos es muy buena y casi idénticas, según los criterios para la selección de modelos, el modelo de RLCPC es mejor que el modelo de RLB.

Se recomienda incrementar el uso de los análisis de reducción de dimensiones en las investigaciones biomédicas, profundizar en el análisis de la regresión logística binaria de componentes principales categóricos, probar el análisis de RLCPC en el contexto de modelos de regresión logística binaria no parsimoniosos y emplear el ACPC como herramienta para la selección de variables en el análisis de RLB.

REFERENCIAS

- [1] Aygencel G., Turkoglu M., Turkoz Sucak G., Benekli M. (2014): Prognostic factors in critically ill cancer patients admitted to the intensive care unit. **J Crit Care.** 29: 618-26.
- [2] Vincent J.L., Moreno R. (2010): Clinical review: Scoring systems in the critically ill. **Critical Care.** 14: 207.

- [3] Wang S., Jiang X., Wu Y., Cui L., Cheng S., Ohno-Machado L. (2013): Expectation propagation logistic regression (EXPLORER): Distributed Privacy-Preserving Online Model Learning. **J Biomed Inform.** 46 (3): 480-96.
- [4] Gupta H., Gupta P.K., Fang X., Miller W.J., Cemaj S., Forse R.A., et al. (2011): Development and Validation of a Risk Calculator Predicting Postoperative Respiratory Failure. **Chest.** 140 (5): 1207-15.
- [5] Hosmer D.W., Lemeshow S., Sturdivant R.X. (2013): **Applied Logistic Regression.** 3rd edition. New Jersey: John Wiley and Sons.
- [6] Linnet K., Bossuyt P.M.M., Moons K.G.M., Reitsma J.B. (2012): Quantifying the Accuracy of a Diagnostic Test or Marker. **Clinical Chemistry.** 58 (9): 1292-1301.
- [7] Lu M., Lee H.S., Hadley D., Huang J.Z., Qian X. (2014): Supervised categorical principal component analysis for genome-wide association analyses. **BMC Genomics.** 15 (Suppl. 1): S10.
- [8] Vavougiou G.D., Natsios G., Pastaka C., Zarogiannis S.G., Gourgoulidis K.I. (2015): Phenotypes of comorbidity in OSAS patients: combining categorical principal component analysis with cluster analysis. **J Sleep Res.** Sep 14. Doi: 10.1111/jsr.12344. [Epub ahead of print].
- [9] de Leeuw J. (2011): **Nonlinear Principal Component Analysis** [Internet]. 2011 Oct. 25. Department of Statistics, UCLA. Disponible en: <https://escholarship.org/uc/item/15j4b16j>.
- [10] Jolliffe I.T., Cadima J. (2016): Principal component analysis: a review and recent developments. **Phil Trans R Soc A.** 374: 20150202. Doi: <http://dx.doi.org/10.1098/rsta.2015.0202>.
- [11] Budaev S.V. (2010): Using Principal Components and Factor Analysis in Animal Behaviour Research: Caveats and Guidelines. **Ethology.** 116 (5): 472-80.
- [12] Wang X., O'Dwyer N., Halaki M. (2013): A review on the coordinative structure of human walking and the application of principal component analysis. **Neural Regen Res.** 8 (7): 662-70.
- [13] Chen X., Wang L., Hu B. (2010): Pathway-based analysis for genome-wide association studies using supervised principal components. **Genetic Epidemiol.** 34: 716-24.
- [14] Ma S., Kosorok M.R. (2009): Identification of differential gene pathways with principal component analysis. **Bioinformatics.** 25: 882-9.
- [15] Wu Y., Jiang X., Kim J., Ohno-Machado L. (2012): Grid binary logistic regression (GLORE): building shared models without sharing data. **J Am Med Inform Assoc.** 19: 758-64.
- [16] Hosmer D.W., Hosmer T., Le Cessie S., Lemeshow S. (1997): A comparison of goodness-of-fit tests for the logistic regression model. **Stat Med.** 16: 965-980.
- [17] Ware J.H. (2006): The limitations of risk factors as prognostic tools. **N Engl J Med.** 355: 2615-7.
- [18] Han K., Song K., Choi B.W. (2016): How to Develop, Validate, and Compare Clinical Prediction Models Involving Radiological Parameters: Study Design and Statistical Methods. **Korean J Radiol.** 17 (3): 339-50.
- [19] Kampfrath T., Levinson S.S. (2013): Brief critical review: Statistical assessment of biomarker performance. **Clin Chim Acta.** 419: 102-7.

UN ANÁLISIS MATEMÁTICO DE EPIDEMIAS TRANSMITIDAS POR EL AEDES AEGYPTI: DENGUE, CHIKUNGUNYA Y ZIKA.

Carlos Bouza*, Sira Allende* José M. Sautto** & Malay Taylor***

*Universidad de La Habana, La Habana. Cuba

**Univerisdad Automa de Guerrero, Acapulco. Mexico

***Taylor Medical Institute, Kolkata. India

ABSTRACT

In this paper we develop and evaluate a model for compartmenting in classes humans and mosquitoes: susceptible, exposed, infected and recovered-immune. A set of differential equations are settled and a simulating experiment is developed for illustration.

KEYWORDS: transmission model, differential equation system, sensitive analysis, simulation

RESUMEN

En este trabajo elaboramos y evaluamos un modelo en el que se compartimentan humanos y mosquitos en clases: susceptibles, expuestos, infectados y recuperados-inmunes. Se determina un conjunto de ecuaciones diferenciales y se desarrolla un experimento de simulación para ilustrar.

1. INTRODUCCIÓN

El mosquito es una fuente de epidemias en todo el mundo. Enfermedades bien conocidas como el Dengue, la Malaria y la Fiebre Amarilla han sido, y son, causa de epidemias serias en las regiones tropicales y subtropicales. Esto hace que estas enfermedades constituyan un interés para los organismos de salud pública en países tan distantes, pero de estas regiones, como Cuba e India.

Es la hembra la que puede hacer la transmisión al picar a los humanos. Estas enfermedades se diseminan grandemente a partir del intercambio creciente de viajeros en todo el mundo. Los cambios climáticos hacen prever que serán hospederos del mosquito zonas templadas en un periodo cercano, ver Semenza-B. Menne (2009) y Massad et al. (2011). La epidemia del Chikungunya está en desarrollo actualmente y comienza a afectar, en forma similar al Dengue en todo el planeta, dadas las similitudes entre ellas, vea Schwartz-Albert (2010). Ambas son transmitidas, en lo fundamental, por una misma especie de mosquito. Para ambas el vector fundamental son los mosquitos *Aedes aegypti* y *Aedes albopictus*. Contraer estas tiene un efecto inmunizador en los humanos.

La medicina ataca estas epidemias tratando fundamentalmente de cortar la trasmisión. Las políticas usuales son la fumigación de las áreas afectadas, la eliminación de hospederos del mosquito y el aislamiento de los infectados. Sin embargo, la magnitud del problema deja muchos aspectos esenciales fuera del control y estudio. El establecimiento de aspectos esenciales y la relación entre diversos aspectos, queda a la interpretación de los especialistas. Lo complicado del problema deja abierta interrogantes sobre el posible efecto de todas las políticas sanitarias alternativas sobre las epidemias y la importancia de ciertos parámetros.

Los modelos matemáticos pueden ser importantes para evaluar los efectos de políticas sanitarias, sin comprometer el desarrollo de la lucha contra las epidemias en un tiempo y lugar concretos. Estos modelos van a permitir comparar los impactos de las decisiones al largo plazo, evaluando mediante simulación el desarrollo de las epidemias. Puede consultarse Dumont et al. (2008), Hartley et al. (2002) y para sus antecedentes vease Mangel-Clark (1988), Nishiura (2006), Yakob-Clements, A.C., (2013), por ejemplo.

En este trabajo elaboramos y evaluamos un modelo en el que se compartimentan humanos y mosquitos en clases: susceptibles, expuestos, infectados y recuperados-inmunes. Cuando un mosquito infectado pica a un humano susceptible, este adquiere la enfermedad con una cierta probabilidad (tasa) pasando a la clase de los expuestos y de esta a la de los infectados, si adquiere tras el periodo de incubación la enfermedad, con otra tasa. Al recuperarse el humano pasa a la clase de los recuperados. Estando infectado un

humano puede contaminar a un mosquito de la clase de los expuestos. Los humanos pueden salir de la población por su deceso, lo que se efectúa a partir una cierta tasa.

La sección 2 se dedica a presentar algunos elementos de la historia del Dengue la Chikungunya y el Zika. También se caracteriza la dinámica de transmisión

La sección 3 presenta los más comúnmente utilizados modelos para la transmisión.

La sección 4 es donde se desarrolla la discusión sobre los modelos matemáticos a utilizar.

En la quinta sección se hace un análisis de los resultados del modelo propuesto en base a parámetros dados en la literatura.

2. LAS ENFERMEDADES BAJO MODELACIÓN

Los mosquitos transmiten una variedad de los llamados arbovirus (arthropod-borne viruses). El mosquito *Aedes aegypti* ha mostrado su capacidad para adaptarse a cambios ambientales tanto naturales como incididos por el hombre.

Ha sido probado algebraicamente que, si se mantiene un nivel de control mínimo, pero constante, se puede garantizar que su número básico reproductivo se mantenga por debajo de uno. Esto garantiza que los índices que establecen un nivel biológicamente realista sean aceptable

Consideraremos los arbovirus transmitidos por el *Ae. aegypti* and *Ae. albopictus*, que es el principal vector del dengue en sus variedades DENV-1, DENV-2, DENV-3 y DENV-4, el Chikungunya (CHIKV), la fiebre amarilla (YFV) y la más reciente epidemia del Zika (ZIKV). Este mosquito es original del África y posee una probada capacidad vectorial dada por su comprobada efectividad de transmitir virus a través de la saliva. Esto está bien documentado para el DENV, CHIKV, ZIKV, and YFV. Los interesados en ampliar pueden consultar Chowell et al. (2007), Hartley et al. (2002) y Nishiura, (2006):

Actualmente tanto el Dengue como el Chikungunya se han desarrollado grandemente. Esto es debido a la expansión de los intercambios entre colectivos humanos de diversos países y continentes. Estos tienen características similares y hay aspectos que les hacen distinguibles.

Una epidemia de Dengue tuvo lugar simultáneamente en 1780, teniendo como escenarios países de Asia, África and Norte América. Esta enfermedad fue identificada en 1779. Su lugar natural de expansión son las regiones del tipo tropical y subtropical. Su expansión geográfica actual se atribuye a la coexistencia de cuatro cepas del virus del dengue y sus vectores.

Según la World Health Organization cada año hay entre 50 y 100 millones de humanos infectados con Dengue. De ellos, medio millón son aquejados del llamado Dengue hemorrágico. Se calcula en unas 22,000 el número de muertes anuales. Estas se reportan generalmente en niños. Vea amplia discusión de estos aspectos en WHO (2005, 2012 y 2016)

El Chikungunya por su parte es una forma relativamente rara de fiebre viral, causada por un alfa virus y su vector principal también es el mosquito *Aedes aegypti*. Esta enfermedad fue vez descrita por vez primera en Tanzania, África en 1952. Este nombre proviene de la palabra Makonde cuyo significado es "el que se dobla". Esto se asocia a la frecuente postura de los enfermos al ser aquejado de síntomas artríticos. Los vectores identificados hasta ahora son los mosquitos de las especies *Aedes*, *Culex* y *Mansonia*. Esta enfermedad se asocia a posibles complicaciones neurológicas como la meningoencefalitis. Además, aunque raramente, se puede transmitir de madre embarazada a su hijo el virus. Vea detalles en Dumont et al. (2008), Paupy et al. (2012), Pialoux et al. (2007):

El Zika por su parte posee síntomas similares a las formas leves del dengue. Se ha ido extendiendo a través de océano Pacífico. Actualmente ha pasado a poseer niveles pandémicos. Es probada la transmisión de mujeres embarazadas a sus hijos. Entre sus efectos esta la microcefalia en los bebés y otros problemas cerebrales. Se reportan cerca de 5,000 casos de microcefalia en bebés nacidos de madres con Zika. Esta cifra es considerada una sub-estimación de la real. En los adultos puede conllevar el síndrome de Guillain-Barré. El virus puede transmitirse sexualmente y por transfusiones de sangre. Para ampliar vea Gatherer- Alain (2015).

La World Health Organization (WHO) considera en sus documentos que la epidemia del Zika genera un problema de emergencia internacional. El tratamiento del Zika debe ser similar pero no tenemos experiencia acumulada suficiente sobre ella. Se incluye esta en el estudio considerando despreciable el número de madres embarazadas. Más elementos clínicos puede obtenerse en fuentes como Gatherer, D. y A. Kohl, Alain (2015).

Los problemas de modelación pueden verse con más detalle en Dumont et al. (2008),

3. TIPOS DE MODELOS PARA LA TRANSMISIÓN

Consideremos que se cumplen las condiciones siguientes:

1. Todo individuo de la población tiene la misma probabilidad de contagiarse con una tasa β .
2. Se satisface la Ley de la Acción de Masa.

Esta segunda condición establece que en el proceso simultáneamente ocurre lo siguiente:

- a) Los individuos que salen de la clase de los susceptibles pasan a la de infectados.
- b) Los que salen de la clase de los infectados pasan a la de los removidos a partir de una fracción γ .

La tercera condición es

3. La tasa de infestación y recuperación es mucho más rápida que la de nacimientos y muertes

Esta condición permite no considerar en el modelo cambios en las clases por factores externos a la epidemia.

El tratamiento clásico parte de considerar que la población en un instante t es de tamaño

$$N = S + I + R$$

Representemos el su flujo como el dado en la figura 3.1.

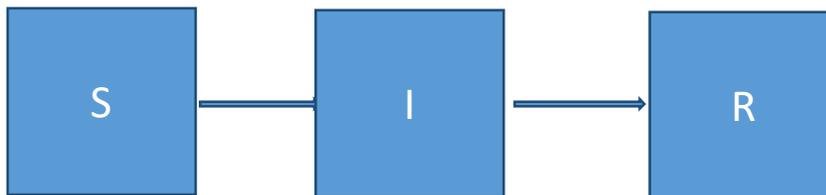


Figura 3.1. SIR

Este modelo es denominado SIR. Podemos reinterpretar las variables envueltas definidas en la Tabla 3.1 como sigue:

Tabla 3.1. SIR

	Número de individuos Susceptibles no infectados o los susceptibles enfermos
	Número de individuos infectados o capaces de contaminar la enfermedad a los susceptibles.
	Conjunto de los infectados ya recuperados (no pueden transmitir la enfermedad)

Note que la primera condición establece que un infectado al hacer contacto es capaz de transmitir la enfermedad a $N(I)=\beta N$ individuos de la población, Así, como la fracción de contaminables es S/N , el número de infectados por unidad será $\beta N \times \left(\frac{S}{N}\right)$. Los nuevos infectados abandonan el grupo de los susceptibles. En realidad tendremos $\beta N \times \left(\frac{S}{N}\right) \times I = \beta SI$. Entonces tendremos la ecuación diferencial

$$\frac{dS}{dt} = -\beta SI$$

La segunda condición nos sugiere fijar las ecuaciones diferenciales

$$\frac{dI}{dt} = \beta SI - \gamma I$$

$$\frac{dR}{dt} = \gamma I$$

Si incluyéramos estos factores el sistema de ecuaciones diferenciales para el modelo SIR sería

Modelo SIR

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI + M(N - S) \\ \frac{dI}{dt} &= \beta SI - \gamma I - MI \\ \frac{dR}{dt} &= \gamma I - MR\end{aligned}$$

Una variante del modelo SIR es considerar que los infectados al recuperarse no son inmunes. Entonces el flujo es

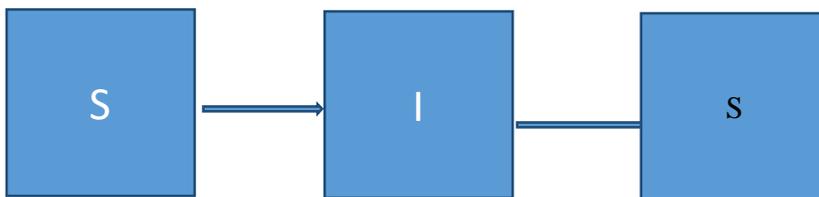


Figura 3.2. SIS

Este es llamado SIS y es representado por el sistema de ecuaciones diferenciales siguiente:

Modelo SIS

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI + M(N - S) + \gamma I \\ \frac{dI}{dt} &= \beta SI - \gamma I - MI - \gamma I\end{aligned}$$

Si la resistencia a la enfermedad tras recuperarse es solo temporal tendremos el modelo SIRS

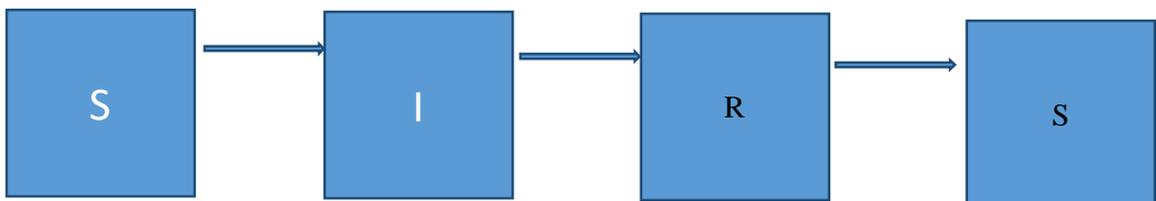


Figura 3.3. SIRS

Este es modelado por el sistema

Modelo SIRS

$$\begin{aligned}\frac{dS}{dt} &= -\beta SI + M(N - S) + fR \\ \frac{dI}{dt} &= \beta SI - \gamma I - MI \\ \frac{dR}{dt} &= \gamma I - MR - fR\end{aligned}$$

Considerando los periodos de exposición y latencia de la enfermedad tendremos la existencia de la clase bajo este escenario tenemos modelos derivados de SIR y SIS.

Del SIR se deriva el SEIR

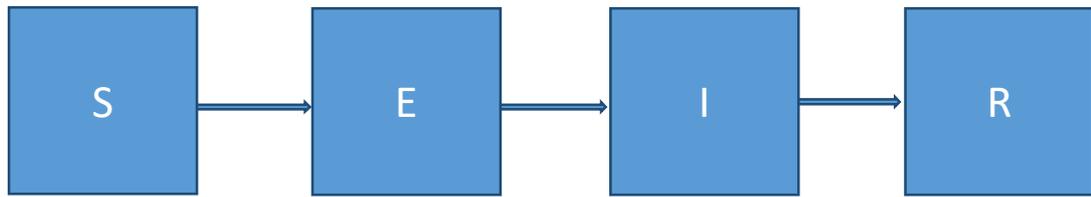


Figura 3. SEIR

Este es descrito por

Modelo SEIR

$$\begin{aligned} \frac{dS}{dt} &= B - \beta SI - MS \\ \frac{dE}{dt} &= \beta SI - (\varepsilon + M)E \\ \frac{dI}{dt} &= \gamma I - MR \end{aligned}$$

El modelo SIS se extiende al SEIS el es esquematizado en la siguiente figura

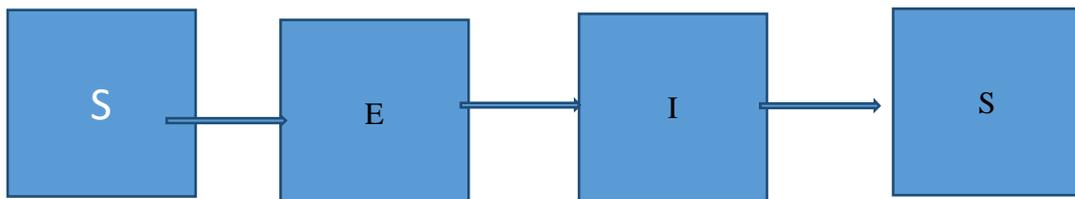


Figura 3.5. SEIR

Su sistema de ecuaciones diferenciales es

Modelo SEIS

$$\begin{aligned} \frac{dS}{dt} &= B - \beta SI - MS + \gamma I \\ \frac{dE}{dt} &= \beta SI - (\varepsilon + M)E \\ \frac{dI}{dt} &= \varepsilon E - (\gamma + M)I \end{aligned}$$

4. LA MODELACIÓN MATEMÁTICA

4.1. La modelación

Los modelos matemáticos pueden ser de ayuda al establecer relaciones no transparentes del estudio de la efectividad de políticas de control y para entender la evolución de mosquitos, virus y transmisión. Para ello debe fijarse cómo interactúan los parámetros médicos de la epidemia y establecer su importancia a través de su rol en los modelos y la sensibilidad de estos a cambios. Los cambios se suponen implementados a través de políticas de control. Reducir el fenómeno a un modelo matemático es una simplificación de la realidad pero provee de una mirada hacia la esencia de la fenomenología asociada a las epidemias. Ejemplos de modelación son presentados en Otero-Schweigmann-Solaris (2008), Rodrigues-Monteiro-Torres (2010): entre otros. El modelo que consideraremos establece la división en clases de humanos y mosquitos. Este es el enfoque común desde los inicios de la modelación de epidemias donde los mosquitos son el vector transmisor principal. Vea por ejemplo a Bailey. (1987), Hethcote (2000), Shah-Gupta (2014), Nuraini et al. (2007):

En estas enfermedades los individuos son susceptibles al ser expuestos y pueden ser infestados. Los infectados se recuperan y son inmunes para pasar, tras un tiempo a la población de susceptibles de nuevo. Por tanto es un modelo del tipo SEIRS. En estas enfermedades hay dos poblaciones. La de humanos $H = \{1, \dots, N_H\}$ y la de mosquitos $Q = \{1, \dots, N_Q\}$. Veamos los parámetros y clases que aparecen al modelar nuestro caso

	V=H, Humanos	V=Q, Mosquitos
S v	Número de humanos susceptibles	Número de mosquitos susceptibles
E v	Número de humanos expuestos	Número de mosquitos expuestos
I v	Número de humanos infestados	Número de mosquitos infestados
N v	Número de humanos	Número de mosquitos
R v	Número de humanos recuperados	-----
β vv*	Tasa de contacto Humanos mosquitos	Tasa de contacto mosquitos humanos
Ψ v	Tasa de nacimientos per cápita de humanos	Tasa de nacimientos per cápita de mosquitos
P v	Tasa de a progresión de humanos que pasan de expuestos a infectados	Tasa de a progresión de mosquitos que pasan de expuestos a infectados
γ v	Tasa de recuperación per cápita de humanos que pasan de infectados a recuperados	----
δ v	Tasa de muerte per cápita inducida por la enfermedad de humanos	---
ρ v	Tasa de perdida dela inmunidad per cápita de humanos	---
λ v	Tasa con que se infestan los humanos por unidad de tiempo	Tasa con que se infestan los mosquitos por unidad de tiempo
μ v1	Parte de la densidad independiente de la tasa de la muerte de humanos	Parte de la densidad independiente de la tasa de la muerte de mosquitos
μ v2	Parte de la densidad dependiente de la tasa de la muerte de humanos	Parte de la densidad dependiente de la tasa de la muerte de mosquitos
η v	---	Tasa de picadas de los mosquitos per cápita

Podemos fijar un sistema de ecuaciones diferenciales que describe el modelo de la epidemia. Las poblaciones son de tamaños respectivos

$$N_H = S_H + E_H + I_H + R_H$$

$$N_Q = S_Q + E_Q + I_Q$$

Así, la diseminación de las enfermedades transmisible por el mosquito puede ser estudiado tomando dos sistemas de ecuaciones diferenciales.

A partir de la variables de estados y parámetros para la Chikungunya y el Dengue satisfacen las ecuaciones

Para Humanos

Sistema H

$$\frac{dS_H}{dt} = L_H + \Psi_H N_H + \rho_H R_H - \lambda_H S_H - (\mu_{H1} + \mu_{H2} N_H) S_H$$

$$\frac{dE_H}{dt} = \lambda_H S_H - P_H E_H - (\mu_{H1} + \mu_{H2} N_H) E_H$$

$$\frac{dI_H}{dt} = P_H E_H - \lambda_H I_H - (\mu_{H1} + \mu_{H2} N_H) I_H - \delta_H I_H$$

$$\frac{dR_H}{dt} = \gamma_H I_H - \rho_H R_H - (\mu_{H1} + \mu_{H2} N_H) R_H$$

Para el análisis es más conveniente llevar a reales todos los valores envueltos.

$$e_H = \frac{E_H}{N_H}; \quad i_H = \frac{I_H}{N_H}; \quad r_H = \frac{R_H}{N_H}; \quad s_H = \frac{S_H}{N_H}$$

De acá se deriva la ecuación $S_H = s_H N_H = (1 - e_H - i_H - r_H) N_H$

Entonces tenemos

Sistema H1

$$\begin{aligned} \frac{dE_H}{dt} &= \frac{de_H}{dt} N_H + \frac{e_H dN_H}{dt} \\ \frac{dI_H}{dt} &= \frac{di_H}{dt} N_H + \frac{i_H dN_H}{dt} \\ \frac{dR_H}{dt} &= \frac{dr_H}{dt} N_H + \frac{r_H dN_H}{dt} \end{aligned}$$

Para mosquitos

Sistema Q

$$\begin{aligned} \frac{dS_Q}{dt} &= \psi_Q N_Q - \lambda_Q S_Q - (\mu_{Q1} + \mu_{Q2} N_Q) S_Q \\ \frac{dE_Q}{dt} &= \lambda_Q S_Q - P_Q E_Q - (\mu_{Q1} + \mu_{Q2} N_Q) E_Q \\ \frac{dI_Q}{dt} &= P_Q E_Q - (\mu_{Q1} + \mu_{Q2} N_Q) I_Q \end{aligned}$$

Los valores fraccionarios son en este caso

$$e_Q = \frac{E_Q}{N_Q}; \quad i_H = \frac{I_Q}{N_Q}; \quad s_Q = \frac{S_Q}{N_Q}$$

Para los mosquitos tenemos

Sistema Q1

$$\begin{aligned} \frac{dE_Q}{dt} &= \frac{de_Q}{dt} N_Q + \frac{e_Q dN_Q}{dt} \\ \frac{dI_Q}{dt} &= \frac{di_Q}{dt} N_Q + \frac{i_Q dN_Q}{dt} \end{aligned}$$

Así que tendríamos la ecuación $S_Q = s_Q N_Q = (1 - e_Q - i_Q) N_Q$.

Es relativamente sencillo resolver los sistemas H1 y Q1, en términos de las variables fraccionarias (reales). Sustituyéndolas en los sistema H y Q

Sumando las ecuaciones iniciales de los humanos tenemos que

$$\frac{dN_H}{dt} = L_H + \psi_H N_H - (\mu_{H1} + \mu_{H2} N_H) N_H - \delta_H I_H$$

Mientras que para los mosquitos

$$\frac{dN_Q}{dt} = \psi_Q N_Q - (\mu_{Q1} + \mu_{Q2} N_Q) N_Q$$

Sea β_{QH} la tasa a contactos entre mosquitos y humanos. Recuerde que σ_Q es la tasa de picadas de los mosquitos. De las sustituciones en H y Q de los resultados de resolver H1 y Q1 obtenemos un nuevo sistema de ecuaciones

$$\frac{de_H}{dt} = \frac{\sigma_Q \beta_{QH} N_Q}{N_H} i_H (1 - e_H - i_H - r_H) - \left(P_H + \psi_H + \frac{L_H}{N_H} \right) e_H + \delta_H e_H i_H$$

$$\frac{di_H}{dt} = P_H e_H - \left(\gamma_H + \psi_H + \delta_H + \frac{L_H}{N_H} \right) i_H + \delta_H i_H^2$$

$$\frac{dr_H}{dt} = \gamma_H i_H - \left(\psi_H + \rho_H + \frac{L_H}{N_H} \right) r_H + \delta_H i_H r_H$$

Se tiene que

$$\frac{dN_H}{dt} = L_H + \psi_H N_H - (\mu_{H1} + \mu_{H2} N_H) N_H - \delta_H I_H$$

$$\frac{dN_Q}{dt} = \psi_Q N_Q - (\mu_{Q1} + \mu_{Q2} N_Q) N_Q$$

$$\frac{de_Q}{dt} = \sigma_Q(\beta_{QH}i_H + \bar{\beta}_{QH}r_H)(1 - e_Q - i_Q) - (P_Q + \Psi_H)e_Q$$

$$\frac{di_Q}{dt} = P_Q e_Q - \Psi_Q i_Q$$

$\bar{\beta}_{QH}$ representa el promedio de la tasa de contactos de mosquitos y humanos.

Es claro que el tamaño de las poblaciones son positivos ($N_H > 0$ y $N_Q \geq 0$).

4.2. Elementos para la simulación de la epidemia.

Si no hay enfermos tenemos que $e_H = e_Q = i_H = i_Q = r_H = 0$. El significado de esto es que N_H es la raíz positiva de una ecuación de segundo grado. Esta es:

$$N_H = \frac{\Psi_H - \mu_{H1} + \sqrt{(\Psi_H - \mu_{H1})^2 + 4\mu_{H2}L_H}}{2\mu_{H2}}$$

Por otra parte

$$N_Q = \frac{\Psi_Q - \mu_{Q1}}{\mu_{Q2}}$$

Entonces si la enfermedad no está a manifestada tenemos el punto

$$\left(e_H, i_H, r_H, \frac{\Psi_H - \mu_{H1} + \sqrt{(\Psi_H - \mu_{H1})^2 + 4\mu_{H2}L_H}}{2\mu_{H2}}, e_Q, i_Q, \frac{\Psi_Q - \mu_{Q1}}{\mu_{Q2}} \right)$$

Tenemos dos soluciones en tal caso:

Solución 1.

$$\left(0, 0, 0, \frac{\Psi_H - \mu_{H1} + \sqrt{(\Psi_H - \mu_{H1})^2 + 4\mu_{H2}L_H}}{2\mu_{H2}}, 0, 0, \frac{\Psi_Q - \mu_{Q1}}{\mu_{Q2}} \right)$$

Esta representa el hecho de que no hay enfermos en la población.

Solución 2.

Esta fija el hecho de que no hay mosquitos.

$$\left(0, 0, 0, \frac{\Psi_H - \mu_{H1} + \sqrt{(\Psi_H - \mu_{H1})^2 + 4\mu_{H2}L_H}}{2\mu_{H2}}, 0, 0, 0 \right)$$

Un modelo usual es considerar que la abundancia de mosquitos es una variable con distribución de Poisson. Así que la probabilidad de que haya al menos un mosquito en un cierto lugar es $P(X \geq 1) = 1 - e^{-\lambda}$, donde λ es abundancia esperada de mosquitos. Entonces se puede estimar esta esperanza.

La matriz para la próxima generación de mosquitos es

$$NGM = \begin{bmatrix} 0 & \frac{p_H \sigma_{QH}}{\mu_Q} \\ \sigma_{QH} p_Q \left(\frac{N_Q}{N_H(\gamma + \mu_Q)} \right) & 0 \end{bmatrix} = \begin{bmatrix} 0 & K_{HQ} \\ K_{QH} & 0 \end{bmatrix}$$

Podemos apuntar que :

- $\frac{p_H \sigma_{QH}}{\mu_Q}$ es el factor de generación de la transmisión del mosquito al humano
- $\sigma_{QH} p_Q \left(\frac{N_Q}{N_H(\gamma + \mu_Q)} \right)$ es el factor de generación de la transmisión del humano a mosquito

NGM tiene dos raíces

$$\mp \sqrt{\frac{p_H \sigma_{QH}}{\mu_Q} \times \sigma_{QH} p_Q \left(\frac{N_Q}{N_H(\gamma + \mu_Q)} \right)}$$

Así que la tasa básica de reproducción es el valor propio mayor de NGM, o sea

$$R_0 = \sqrt{\frac{p_H p_Q \sigma_{QH}^2}{\mu_Q}} \times \left(\frac{N_Q}{N_H(\gamma + \mu_Q)} \right)$$

Este juega un papel importante en el estudio de las epidemias ver por ejemplo Bacaër (2007): En diversos trabajos este índice es estimado a partir de considerar para cada instante t la validez de la relación $K(t) \propto e^{\lambda(t)}$

(t) es el número de casos de dengue acumulados y (t) es denominada fuerza de la infección.

Para estimar R_0 se considera que en el minuto $t=0$ el número de humanos infectados I_H y el de mosquitos I_Q crecen exponencialmente con la misma tasa en un periodo corto de tiempo ($t \approx 0$).

Otros enfoques hacen uso de regresión logística para establecer el efecto de covariables como edad, sexo etc. Al hacer uso de estos modelos es posible establecer, usando estimación del tipo cuasi-máximo verosímil, la validez de usar una distribución como la Poisson u otra que pueda describir el fenómeno como la Binomial negativa.

Sobre los diferentes problemas de la estimación de R_0 se recomienda consultar Jafaruddin et al. (2015), Allen- van den Driessche, (2008),.

5. ESTUDIO DE LA DESCRIPCIÓN DADA POR EL MODELO.

Usaremos como base valorar de los parámetros existentes en la literatura, vea XXX, para el Dengue (DENV) y la Chikungunya (CHIK-A). A partir de ellos determinamos valores que son aceptables para ambas epidemias. Estos se han reportado en diversos papers como Michael et al. Bundy (2001), Roitberga, B. D. y M. Mangel (2010), Van den Driessche,-Watmough (2002).

Tabla 5.1 Datos Bio-epidemiológicos de humanos y mosquitos.

	V=H, Humanos	V=Q, Mosquitos	
S_V	Número de humanos susceptibles	Número de mosquitos susceptibles	0- 7,96 S_H
β_{VV^*}	Tasa de contacto Humanos mosquitos 0,022	Tasa de contacto mosquitos humanos 0,24	
E_V	Número de humanos expuestos	Número de mosquitos expuestos	
I_V	Número de humanos infestados	Número de mosquitos infestados	
N_V	Número de humanos	Número de mosquitos	0-2 N_H
R_V	Número de humanos recuperados	----	
L_V	Tasa de emigración de humanos	----	0,088
Ψ_V	Tasa de nacimientos per cápita de humanos $5,6 \times 10^{-5}$	Tasa de nacimientos per cápita de mosquitos 0,13	
P_V	Tasa de a progresión de humanos que pasan de expuestos a infectados	Tasa de a progresión de mosquitos que pasan de expuestos a infectados	0,227
γ_V	Tasa de recuperación per cápita de humanos que pasan de infectados a recuperados	----	0,003-0,1
δ_V	Tasa de muerte per cápita inducida por la enfermedad de humanos	---	0,007
ρ_V	Tasa de pérdida de la inmunidad per cápita de humanos	---	
λ_V	Tasa con que se infestan los humanos por unidad de tiempo 0,0068	Tasa con que se infestan los mosquitos por unidad de tiempo	0,25
μ_{V1}	Parte de la densidad independiente de la tasa de la muerte de humanos $8,8 \times 10^{-6}$	Parte de la densidad independiente de la tasa de la muerte de mosquitos	0,033
μ_{V2}	Parte de la densidad dependiente de la tasa de la muerte de humanos $2,03 \times 10^{-5}$	Parte de la densidad dependiente de la tasa de la muerte de mosquitos 4×10^{-5}	
σ_V	---	Tasa de picadas de los mosquitos per cápita	0,25-1

....

5.1. Análisis de sensibilidad

Para determinar la mejor manera de reducir la incidencia de las epidemias en la población de humanos, es necesario conocer la importancia relativa de los diferentes factores responsables de su transmisión.

A partir de la solución del modelo matemático usando los valores dados en la tabla podemos hacer un análisis de sensibilidad. La forma más simple de analizar la sensibilidad consiste en considerar las derivadas de las cantidades de interés como función de los parámetros de los modelos. Esto es usando

$$\gamma_{\theta}^Z = \frac{\partial Z}{\partial \theta} \times \frac{\theta}{Z}$$

Estos índices cuantifican como, pequeños cambios en los valores de entrada, causan una variabilidad en la salida. El signo indica la dirección de la repuesta a los cambios y su valor su importancia relativa para las predicciones. Así se determina la importancia relativa de ellos en la predicción, ver Chitnis et al. (2008): para una detallada discusión. Para hacer los análisis se introducen perturbaciones en los datos d entrada y se mide su efecto a través de la sensibilidad.

La transmisión de la enfermedad está directamente relacionada con R_0 . Por tal razón, se calcularon los índices de sensibilidad del número reproductivo básico, R_0 , para los parámetros del modelo

5.2. Los resultados de la simulación

Tabla 5.2 Sensibilidad de R respecto a los parámetros usados

	V=H, Humanos	
Ψ_Q	Tasa de nacimientos per cápita de mosquitos	+8,9728
σ_Q		+1,0027
β_{HQ^*}	Tasa de contacto Humanos mosquitos	+0,4729
L_H	Tasa de emigración de humanos	+0,3260
P_H	Tasa de a progresión de humanos que pasan de expuestos a infectados	+0,2995
μ_{H1}	Parte de la densidad independiente de la tasa de la muerte de humanos $8,8 \times 10^{-6}$	+0,0186
P_Q	Tasa de a progresión de humanos que pasan de expuestos a infectados	+0,0011
Ψ_H	Tasa de nacimientos per cápita de humanos $5,6 \times 10^{-5}$	-0,0038
γ_H	Tasa de recuperación per cápita de humanos que pasan de infectados a recuperados	-0,0984
μ_{H2}	Parte de la densidad dependiente de la tasa de la muerte de humanos $2,03 \times 10^{-5}$	-0,2618
ρ_H	Tasa de perdida de la inmunidad per cápita de humanos	-0,2649
δ_H	Tasa de muerte per cápita inducida por la enfermedad de humanos	-0,2816
μ_{Q2}	Parte de la densidad de mosquitos dependiente de la tasa de la muerte de humanos	-0,4978
μ_{Q1}	Parte de la densidad independiente de la tasa de la muerte de humanos	-7,884

....

Del análisis de sensibilidad se desprende que:

1. Los parámetros más sensibles son la Tasa de nacimientos per cápita de mosquitos (Ψ_Q), la tasa de picadas de mosquitos (σ_Q) y la tasa de contactos entre humanos y mosquitos (β_{HQ})
2. La Tasa de emigración de humanos es también muy importante.
3. La parte de la densidad de mosquitos en la muestra opera inversamente en forma muy significativa.

De estos resultados se desprende que lo más importante es atacar el crecimiento de la población de mosquitos, disminuir la tasa de contactos entre ambas poblaciones. Aparece como conveniente el detener los flujos de humanos.

6. LA POLÍTICA PREVENTIVA EN CUBA

Las campañas preventivas en Cuba tienen por objetivo es eliminar y/o disminuir los factores que condicionan la alta infestación de *Aedes aegypti* y *albopictus*. Esta establecido una estricta vigilancia en los puntos de entrada del país. En particular los viajeros procedentes de naciones con transmisión recién una atención especial. Esto junto con las campañas de desinfección y control de las posibles fuentes de cría de

mosquitos apunta a lo que nuestro análisis determina como políticas más recomendables. Puede hacerse una consulta detallada sobre estas campañas en <http://www.acn.cu/salud/16582-anuncia-minsap-medidas-para-enfrentar-situacion-epidemiologica>.

REFERENCIAS

- [1] Allen, L.J. and P. van den Driessche (2008): The basic reproduction number in some discrete-time epidemic models, *J. Difference Equ. Appl.* 14, 1127–1147
- [2] Arriola, L.M., Hyman, J.M. (2007): Being sensitive to uncertainty. *Computing Sci. Eng.* 9, 10–20.
- [3] Bacaër, N. (2007): Approximation of the basic reproduction number R_0 for vector-borne diseases with aperiodic vector population. *Bull. Math. Biol.* 69, 1067–1091.
- [4] Bailey N. T. J. (1987): *The Mathematical Theory of Infectious Diseases*. Oxford University Press, Oxford.
- [5] Chitnis, N., Hyman, J., Cushing, J. (2008): Determining important parameters in the spread of malaria through the sensitivity analysis of a mathematical model. *Bull. Math. Biol.* 70, 1272–1296.
- [6] Chowell, G., P. D. Dueñas, J. C. Miller et al. (2007): Estimation of the reproduction number of dengue fever from spatial epidemic data. *Mathematical Biosciences*, 208, 571–589.
- [7] Dumont, Y., Chiroleu, F., Domerg, C. (2008): On a temporal model for the Chikungunya disease: modeling theory and numerics. *Math. Biosci.* 213, 80–91.
- [8] Gatherer, D. y A. K., Alain (2015): Zika virus: a previously slow pandemic spreads rapidly through the Americas". *Journal of General Virology*, 97, 269–273..
- [9] Hartley, L.M., Donnelly, C.A. y Garnett, G.P. (2002): The seasonal pattern of dengue in endemic areas: mathematical models of mechanisms. *Trans. R. Soc. Trop. Med. Hyg.* 96, 387–397.
- [10] Hethcote, H. W. (2000): The mathematics of infectious diseases, *SIAM Rev.* 42, 599–653.
- [11] Jafaruddin, S., W. Indratno, N. Nuraini, A. K. Supriatna y E. Soewono (2015): Estimation of the Basic Reproductive Ratio for Dengue Fever at the Take-Off Period of Dengue Infection. *Computational and Mathematical Methods in Medicine*, Article ID 206131, 14 pages. <http://dx.doi.org/10.1155/2015/206131>
- [12] Mangel, M. y Clark, C.W. (1988): *Dynamic modeling in behavioral ecology*. Princeton University Press, Princeton, N. Jersey.
- [13] University Press, Princeton, N. Jersey.
- [14] Massad, E., Coutinho, F.A.B., Lopez, L.F., da Silva, D.R. (2011): Modeling the impact of global warming on vector-borne infections. *Phys. Life Rev.* 8, 169–199.
- [15] Michael, E., K. D. Ramaiah, S. L. Hoti, G. Barker, M. R. Paul, J. Yuvaraj, P. K. Das, B. T. Grenfell y D. A. P. Bundy (2001): Quantifying mosquito biting patterns on humans by dna fingerprinting of bloodmeals. *Am. J. Trop. Med. Hyg.*, 65, 722–728
- [16] Nishiura, H. (2006): Mathematical and statistical analyses of the spread of dengue, *Dengue Bull.* 30, 51–67
- [17] Nuraini, N., E. Soewono y K. A. Sidarta (2007): Mathematical model of dengue disease transmission with severe DHF compartment. *Bulletin of the Malaysian Mathematical Sciences Society*, 30, 143–157.
- [18] Otero, M., N. Schweigmann, y H.G. Solaris (2008): A stochastic spatial dynamical model for *Aedes aegypti*, *Bull. Math. Biol.* 70, 1297–1325.
- [19] Paupy, C., KassaKassa, F., Caron, M., Nkoghé, D., Leroy, E.M. (2012): A chikungunya outbreak associated with the vector *Aedes albopictus* in remote villages of Gabon. *Vector-Borne Zoonotic Dis.* 12, 167–169.
- [20] Pialoux, G., Gaüzère, B., Jauréguiberry, S., Strobel, M., (2007): Chikungunya: an epidemic arbovirosis. *Lancet Infect. Dis.* 7(5), 319–327.
- [21] Rodrigues, M.H.S., T.T. Monteiro y D.F.M. Torres (2010): Dynamics of dengue epidemics when using optimal control, *Math. Comput. Model.* 52, 1667–1673.
- [22] Rodrigues, H. S., M. T. Monteiro, D. F. M. Torres y A. Zinober (2011): Dengue disease, basic reproduction number and control. *International Journal of Computer Mathematics*, 1–13
- [23] Roitberga, B. D. y M. Mangel (2010): Mosquito Biting and Movement Rates as an Emergent Community Property and The Implications for Malarial Interventions *Israel Journal of Ecology & Evolution*, 56, 297–312.

- [24] Schwartz, O. y Albert, M. (2010): Biology and pathogenesis of Chikungunya virus. *Nat. Rev. Microbiol.* 8, 491–500.
- [25] Shah, N. H. y J. Gupta (2014): Mathematical modeling of Malaria using differential Equations.. In *Deforestation*, (C. Bouza editor), 115-132 .Nova Sc. Pub., N. York.
- [26] Semenza J. C. y B. Menne (2009): , Climate change and infectious diseases in Europe, *Lancet Infect. Dis.* 9, 365–375.
- [27] Van den Driessche, P. y J. Watmough (2002): Reproduction numbers and sub- threshold endemic equilibria for compartmental models of disease transmission. *Math. Biosci.* 180, 29–48.
- [28] WHO (2005): WHO statement on the first meeting of the International Health Regulations (2005) (IHR 2005) Emergency Committee on Zika virus and observed increase in neurological disorders and neonatal malformations. <http://www.who.int/mediacentre/news/statements/2016/1st-emergency-committee-zika/en/>, última consulta marzo, 2016.
- [29] WHO (2012): Dengue guidelines for diagnosis, treatment, prevention and control 2012. New Edition Sheet, World Health Organization, Geneva, Switzerland, 2012, <http://www.who.int/denguecontrol/resources/en/>, última consulta enero 2016.
- [30] WHO (2016): Dengue. Available at <http://www.who.int/topics/dengue>, última consulta enero 2016.
- [31] Yakob, L. y Clements, A.C., (2013): A mathematical model of chikungunya dynamics and control: the major epidemic on Réunion Island. *PLoS One* 8 (3), e57448

Capítulo 12

pp 141-159

PROCEDIMIENTO PARA LA EVALUACIÓN DEL DESARROLLO SOSTENIBLE TERRITORIAL. APLICACIÓN EN LA PROVINCIA DE VILLA CLARA

Anabel Vega Calcines* & Carlos Pérez Risquet**

*Calle 1ra #121 A (altos) e/. D y E, Rpto. Sta Catalina, Santa Clara, Cuba

**Tristá # 510 e/. Amparo y Danielito, CP 50100, Santa Clara, Cuba

ABSTRACT

This investigation proposes a general procedure, and its support tools, for evaluating sustainable development in a territorial scale with the purpose to bring solid basis to decisions making. Its aim is to estimate, and at the same time represent, the relative grade of sustainable development in a territory. On applying the procedure is obtained the Territorial Sustainable Development Index (IDST, for its abbreviation in Spanish) and an indicators system organized by a thematic framework which allows observing with more details the behavior of indicators that compose the index. A web application was developed to support the applying of the procedure and allows designing the indicators system, calculating the index, observing the evolution of indicators, comparing territories and other facilities. The feasibility of applying the procedure and its support tools was demonstrated through its application in Villa Clara province. It permitted to identify critic points and improvement opportunities for guiding effort to the principal problems.

KEYWORDS: indicators of sustainable development, index of territorial sustainable development

RESUMEN

En la presente investigación se propone un procedimiento general y sus instrumentos de apoyo, para la evaluación del desarrollo sostenible a escala territorial con el fin de aportar bases sólidas a la toma de decisiones. El procedimiento tiene por objetivo estimar, y a la vez representar, el grado relativo de desarrollo sostenible de un territorio. Al aplicarlo se obtiene el Índice de Desarrollo Sostenible Territorial (IDST) y un sistema de indicadores organizado mediante un marco ordenador temático, que permite observar con más detalle el comportamiento de los indicadores que componen el índice. Se desarrolló una aplicación web que sirve de apoyo para la aplicación del procedimiento, permitiendo diseñar el sistema de indicadores, calcular el índice, observar la evolución de los indicadores en el tiempo y comparar territorios, entre otros resultados. La factibilidad de aplicación del procedimiento y los instrumentos de apoyo fue demostrada mediante su aplicación en la provincia de Villa Clara, lo que permitió identificar puntos críticos y oportunidades de mejoras para reorientar esfuerzos hacia los principales problemas.

1 INTRODUCCIÓN

La problemática medioambiental se ha convertido, desde de la década de los 70 del pasado siglo, en una de las principales preocupaciones de políticos, ambientalistas, ecologistas, organizaciones no gubernamentales, instituciones científicas y la sociedad en su conjunto. El origen de dichos problemas está en los sistemas de desarrollo que se han asumido, sustentados en patrones de producción y consumo irracionales, causantes del atraso y la pobreza que azota a gran parte de la humanidad y de graves problemas ambientales que suponen un riesgo para nuestra propia especie.

“Cuando se habla de crisis ecológica no se hace en términos abstractos o altruistas: lo que cuestiona es nuestra supervivencia en el futuro, al menos en unas condiciones que podamos considerar dignas para todos.” [12]

A partir de la celebración de la Conferencia de las Naciones Unidas sobre Medio Ambiente y Desarrollo, en Rio de Janeiro en 1992, la política de desarrollo sostenible es considerada como la meta hacia la cual deben dirigirse todas las naciones.

El concepto de desarrollo sostenible, plantea la búsqueda de progreso económico que no descuide aspectos cualitativos, tales como la calidad de vida y la preservación del medioambiente, sin olvidar el compromiso ético con las generaciones venideras. Así, se considera como un proceso multidimensional que afecta al sistema económico, ecológico y social. [11]

Como resultado de la Cumbre de Rio se aprobó la Agenda 21, documento político a través del cual, la comunidad internacional se comprometió a poner en práctica un programa que contiene las estrategias para

prevenir el deterioro del medio ambiente y establecer las bases para un desarrollo sostenible a escala planetaria en el siglo XXI. [10]

Para conocer si realmente las actividades humanas asociadas al desarrollo de la sociedad, pueden considerarse sostenibles se necesitan indicadores. El Capítulo 40 de la Agenda 21 se refiere a la importancia de crear indicadores de desarrollo sostenible “a fin de aportar bases sólidas al proceso de toma de decisiones en todos los niveles y contribuir a una sustentabilidad autorreglamentada, de los sistemas que integran el medio ambiente”. [28]

Las evaluaciones funcionan mejor cuando son parte de un ciclo continuo de reflexión y acción, en lugar de una actividad aislada. Para guiar políticas y acciones, es necesario que las evaluaciones sean completamente integradas en los procesos regulares de toma de decisiones.

En un proceso de desarrollo sustentable los indicadores no constituyen un fin en sí mismo, sino herramientas que permiten conocer tendencias, el cumplimiento de objetivos y brindan la posibilidad de intervenir oportunamente permitiendo una toma de decisiones fundamentada. [16]

La tarea de elaboración de indicadores de sostenibilidad se hace difícil por la ambigüedad propia del concepto. No obstante, se han elaborado numerosas propuestas tanto de sistemas de indicadores como de metodologías para su definición [6], [5], [13], [26], [3], [23].

Algunos de los marcos utilizados para la evaluación de la sostenibilidad se basan en objetivos concretos, mientras otros implementan un enfoque sistémico [3].

Por otro lado, se han elaborado indicadores complejos para medir la distancia al objetivo de sostenibilidad en términos genéricos, como son la conocida Huella Ecológica, el Índice de Planeta Vivo y el Índice de Bienestar Económicamente Sustentable (IBES).

El trabajo con indicadores de desarrollo sostenible ha sido impulsado por numerosos organismos de investigación y esfuerzos internacionales de cooperación [14]. Las importantes contribuciones técnicas y financieras realizadas demuestran que el desarrollo de indicadores ambientales (y de índices relacionados con la sostenibilidad) se ha convertido en una prioridad internacional.

El Gobierno cubano, y la sociedad en su conjunto, han evidenciado su interés por la protección del medio ambiente en la isla. A partir del triunfo revolucionario de 1959, el pensamiento y la acción ambiental cubana han crecido y se han perfeccionado paralelamente al proyecto de elevar el nivel y la calidad de vida del pueblo cubano. Esto se evidencia a través de los siguientes hitos [20]:

- Ley 33 de Protección del Medio Ambiente y del Uso Racional de los Recursos Naturales de 1981
- Decreto Ley 118 de enero de 1990 “Estructura, Organización y Funcionamiento del Sistema Nacional de Protección del Medio Ambiente y su Órgano Rector”
- Promulgación en 1993 del Artículo 27 de la Constitución de la República de Cuba
- Creación del Ministerio de Ciencia Tecnología y Medio Ambiente (CITMA) en 1994
- Ley 81 del Medio Ambiente en 1997
- Establecimiento de la Estrategia Ambiental Nacional en 1997
- Aprobación de los “*Lineamientos de la política económica y social del Partido y la Revolución*” en 2011

La consolidación del marco legal e institucional favoreció que la estadística en este campo se desarrollara junto al proceso de concientización sobre la importancia de preservar y recuperar el medio, con el objetivo de darle respuesta a la necesidad creciente de contar con información adecuada, para tomar decisiones referentes a la protección y mejora del medio ambiente y para un seguimiento de las mismas en términos de un “desarrollo sostenible”. [20]

El concepto de sostenibilidad tiene diferentes ámbitos de aplicación. Desde una perspectiva sectorial es posible hablar de “industrias sostenibles”, “agricultura sostenible” o “turismo sostenible”, mientras que, desde una perspectiva geográfica, se puede aplicar desde el ámbito local al global. [15]

Desde el ámbito sectorial, la mayoría de las investigaciones realizadas en Cuba, están dirigidas a monitorear indicadores de sostenibilidad turística, dada la importancia del turismo para la economía cubana y la necesidad de proteger los ecosistemas costeros y montañosos de los impactos que genera este sector. Entre las principales aproximaciones metodológicas que se han descrito y aplicado en el diseño de indicadores de desarrollo sostenible para destinos y entidades turísticas se encuentran las realizadas por [25], [23] y [16]. Este último calcula el grado de sustentabilidad de forma sintética.

A escala nacional, se han presentado amplios informes que describen la situación económica, social y medioambiental de Cuba a través de indicadores, como los realizadas por [7], [2], [8], [22], dando respuesta a la demanda de información a organismos internacionales o como parte de proyectos nacionales e

internacionales. De manera general en estos informes se presenta una compilación de indicadores, muchas veces asociados a objetivos y metas, como los que dan seguimiento al cumplimiento de los Objetivos de Desarrollo del Milenio. Sólo uno de estos informes [2] posee un enfoque sistémico y en ningún caso se agregan los indicadores por lo que no nos ofrecen un acercamiento al grado de sustentabilidad en términos genéricos.

A nivel provincial, las delegaciones del CITMA coordinan las acciones para el uso racional de los recursos naturales y la prevención, mitigación y solución de los problemas ambientales, con el objetivo de alcanzar el desarrollo sostenible de los territorios. Sin embargo, a pesar de los esfuerzos realizados en materia de indicadores medioambientales y de desarrollo sostenible a escala nacional y sectorial, las autoridades provinciales no cuentan con sistemas de indicadores y procedimientos de evaluación asociados, que les permitan monitorear en qué medida el proceso de desarrollo de los territorios puede considerarse sostenible para el medio ambiente, la sociedad y la economía, y comprobar así la efectividad de las políticas públicas. Obviamente, tampoco existen herramientas informáticas que soporten el trabajo con indicadores de desarrollo sostenible en las provincias y faciliten la presentación de información relevante tanto para la toma de decisiones como para la población en general.

El objetivo general de este trabajo es, por tanto, desarrollar un procedimiento y sus instrumentos metodológicos de apoyo para integrar indicadores y evaluar el desarrollo sostenible en las provincias cubanas, que facilite la identificación de puntos críticos y oportunidades de mejora en el desarrollo sostenible provincial.

El resto de este trabajo está organizado de la forma siguiente. El epígrafe 2 presenta un estudio de las tendencias actuales en temas relacionados con el desarrollo sostenible, sistemas de indicadores de desarrollo sostenible, índices agregados de sostenibilidad existentes, métodos para la definición de índices agregados y la evaluación de la sostenibilidad en Cuba. En el epígrafe 3 se propone un procedimiento para evaluar el desarrollo sostenible a escala provincial, el cual está estructurado en una secuencia lógica que cuenta con cuatro fases: diagnóstico del territorio, estructuración del sistema de indicadores, cálculo del Índice de Desarrollo Sostenible Territorial y análisis de los resultados. En el epígrafe 4 se muestran los resultados de la aplicación del procedimiento propuesto en la provincia de Villa Clara. Finalmente se exponen las conclusiones y recomendaciones generales derivadas de la investigación realizada.

2 EVALUACIÓN DEL DESARROLLO SOSTENIBLE MEDIANTE INDICADORES

Las evaluaciones de sostenibilidad emergieron como una de las herramientas más útiles para hacer operativo el concepto de desarrollo sostenible. Según [12] un problema epistemológico que surge ineludiblemente es si existe un cuerpo teórico unificado que defina el camino metodológico básico para la elaboración de indicadores de sostenibilidad y medio ambiente; pues aunque numerosos autores señalan requisitos que deberían cumplir los referidos indicadores, estas indicaciones son habitualmente de naturaleza puntual o parcial, sin llegar a instituir un método sólido que totalice y guíe la compleja labor de diseño de indicadores de sostenibilidad y medio ambiente, así como el diseño de sistemas de indicadores.

La definición de los indicadores de sostenibilidad refleja el concepto de desarrollo sostenible que se quiere implementar. Su finalidad es indicar de alguna forma si las actividades humanas, el uso de los recursos naturales o determinadas funciones ambientales, asociadas al desarrollo de la sociedad, pueden considerarse sostenibles de acuerdo a algún criterio de sostenibilidad. Este objetivo, marca la diferencia con respecto a indicadores económicos convencionales (por ejemplo, Producto Interno Bruto, Renta per Cápita) que mistifican las auténticas condiciones de equilibrio y sostenibilidad de los modelos de desarrollo. [5]

Entre los trabajos en este sentido destacan los realizados en el campo de los indicadores y sistemas de información para el Desarrollo Sostenible [18], teoría y métodos de indicadores de Desarrollo Sostenible [4], así como en metodologías para el desarrollo de programas nacionales de indicadores de Desarrollo Sostenible [27]. Más recientemente, otros investigadores como [5], [15], [26], [17] y [3] han desarrollado y aplicado metodologías para la evaluar la sostenibilidad desde distintos enfoques, ámbitos y escalas.

Aunque no todas las evaluaciones contemplan los mismos principios, cualquier proceso de medición o valoración de la sostenibilidad debe guiarse por una visión operativa del desarrollo sostenible, basándose en objetivos claros, sobre la base de un enfoque comprensivo u holístico. [3]

En la literatura no es común encontrar referencias sobre el diseño de aplicaciones informáticas que soporten la evaluación del desarrollo sostenible y la comunicación de los resultados, sino que algunas metodologías se apoyan en hojas de cálculo de *Microsoft Excel* para el cálculo de indicadores complejos y en software estadísticos como el *SPSS for Windows* y el *StathGraphics* para la interpretación de los resultados.

Para ganar en claridad expositiva, los aportes realizados en materia de evaluación de la sostenibilidad se pueden agrupar básicamente en tres grupos: indicadores de sostenibilidad física, indicadores de sostenibilidad integral e índices agregados de sostenibilidad.

2.1. Indicadores de sostenibilidad física

Los indicadores de sostenibilidad física difieren de los meramente medioambientales, en base a que los primeros reflejan no sólo las condiciones y presiones medioambientales, sino también el grado en que ciertas presiones o impactos sobre la Tierra pueden afrontarse a largo plazo sin afectar las estructuras y procesos básicos para la vida. [5]

El interés no radica por tanto en el conocimiento de determinado contaminante sino en relacionarlos en términos de distancias con los objetivos de política definidos acerca de emisiones máximas o capacidad de carga crítica. Pueden expresarse en términos de flujos, tasas de crecimiento o tasa de desviación respecto al umbral/objetivo/estándar.

2.2. Indicadores de sostenibilidad integral

Este grupo se refiere a sistemas de indicadores que desde una visión integradora consideran no únicamente los indicadores ambientales, sino también los referidos a las cuestiones sociales, económicas e institucionales. Un sistema de indicadores organizado de acuerdo a un marco conceptual coherente, adquiere mayor utilidad que si se presenta como un mero listado de elementos. A continuación, se describen algunos marcos ordenadores utilizados en la presentación de sistemas de indicadores de desarrollo sostenible.

2.2.1. Fuerza Motriz-Estado-Respuesta (FMER)

El marco de Fuerza Motriz-Estado-Respuesta fue desarrollado y adoptado por la ONU a partir de 1995, tomando como base el modelo Presión-Estado-Respuesta (diseñado por la OCDE en 1993 para reflejar aspectos puramente ambientales) con la idea de incorporar mejor las connotaciones sociales, económicas e institucionales del desarrollo sostenible. Este marco establece los siguientes grupos de indicadores [24]:

- Indicadores de Fuerza Motriz: representan actividades humanas, procesos y patrones que tienen un impacto en el desarrollo sostenible. Corresponden a desarrollo a nivel de empresas, industrias o sectores económicos, así como a tendencias sociales. Ejemplos: tasa de crecimiento de la población, y la emisión de gases de efecto invernadero.
- Indicadores de Estado: proveen una indicación del estado del desarrollo sostenible, o de un aspecto particular de éste, en cierto momento. Corresponden a indicadores cualitativos o cuantitativos. Por ejemplo: expectativa de escolaridad o concentración de contaminantes en zonas urbanas.
- Indicadores de Respuesta: indican opciones de política y otras respuestas sociales a los cambios en el estado del desarrollo sostenible. Estos indicadores proveen una medida de la disposición y efectividad social en la construcción de respuestas. Incluye legislación, regulaciones, instrumentos económicos, actividades de comunicación. Ejemplos: cobertura de tratamiento de aguas, gasto en disminución de la contaminación.

La ventaja de este marco es que está muy extendido internacionalmente al ser recomendado por la OCDE y Naciones Unidas para realizar comparaciones de indicadores ambientales a nivel internacional. Por otro lado, la crítica sobre linealidad y causalidad se mantiene en este esquema.

El marco de referencia FMER resulta ser poco generalizable y de escasa confiabilidad al no considerar la interdependencia entre los diferentes factores y el carácter de multicausalidad que opera en los fenómenos sociales y ambientales. Estas mismas razones conspiran en contra de saber cuál(es) de las medidas propuestas se consideran como efectivas para mejorar el estado. Por estas razones la Comisión de Desarrollo Sostenible(CDS) decidió abandonar el enfoque FMER. [24]

La Agencia Europea de Medio Ambiente, ha aplicado un modelo similar, el de Fuerza Motriz-Presión-Impacto-Respuesta (FMPIR) que precisa además un grupo de indicadores de impacto y ha sido utilizado para la elaboración de los informes del Medio Ambiente en Europa. [5]

2.2.1. Marcos ordenadores temáticos

Otra forma de ordenar los indicadores es esquematizarlos por tema y subtema, para generar cierta lógica en la comunicación de los resultados. Los marcos temáticos en general están estructurados de acuerdo a las condiciones institucionales de los países y a las prioridades de sostenibilidad en sí. Las dinámicas de medio ambiente y sostenibilidad que son multisectoriales y complejas encuentran en estos esquemas una buena manera de presentar los indicadores. Estos marcos han sido utilizados por numerosos países, como Canadá, Suecia, Nueva Zelanda, España, Estados Unidos, y casi todos los países latinoamericanos y caribeños. [24]

También pueden ser utilizados para monitorear el avance respecto a objetivos concretos de medio ambiente y sostenibilidad que se plantean. El sistema mundial de Objetivos de Desarrollo del Milenio (ODM) se estructura en grandes objetivos con metas más concretas, a las cuales se asocian indicadores con metas cuantitativas y temporales para medir los progresos hacia el logro de los objetivos.

En el contexto del Programa de Trabajo en Indicadores de Desarrollo Sostenible de la CDS, se desarrolló en 2001 un marco ordenador temático en el cual se definen 57 indicadores centrales de desarrollo sostenible (los que resultaron más útiles a los países de la prueba piloto), organizados en 4 dimensiones, 18 temas y 38 subtemas. Este marco tiene la ventaja de que los temas y subtemas pueden reflejar los énfasis nacionales tanto de los problemas como de los desafíos del desarrollo sostenible.

2.3. Índices agregados de sostenibilidad

Existen dos grandes corrientes en la discusión sobre indicadores sintéticos de sostenibilidad. La primera se relaciona con el concepto de “sustentabilidad débil”, según el cual, es posible asignar valores monetarios actualizados a los recursos naturales y a los servicios ambientales de la naturaleza, pudiendo estimar así el desgaste del “Capital Natural” en términos monetarios. En esta corriente se destacan: IBES, Riqueza Verdadera y Ahorro Genuino. La segunda corriente parte del concepto de “sustentabilidad fuerte”; sostiene que es necesario determinar la capacidad del planeta para sostener al conjunto de la economía humana y mantener las funciones ecosistémicas que aseguren la vida en general. En este grupo se puede encontrar: Huella Ecológica, Índice del Planeta Vivo, entre otros. [1]

De todos ellos se pueden rescatar elementos valiosos, tales como su capacidad de sintetizar elementos de la dinámica ecológica, y también de la dinámica económica, ecológica y social. Sin embargo, las medidas sintéticas poseen inconvenientes desde el punto de vista metodológico, derivadas de la heterogeneidad de las variables y la simplificación excesiva. En torno a la sostenibilidad, resulta difícil además justificar la selección de las variables y su importancia, ya que no existe consenso en la comunidad científica para ello. [11]

A menudo se tendía a pensar que los indicadores sintéticos y los sistemas de indicadores eran aproximaciones antagónicas, pero la práctica ha demostrado que pueden ser complementarios, pues siempre que los autores lo prevean, todo trabajo conmensuralista¹ implica la generación de componentes previos que pueden expresarse por separado a través de un sistema, de forma que el usuario los utilice de una u otra forma, de acuerdo a sus necesidades. [24]

No obstante la variedad de índices de sostenibilidad que se han definido, se pueden elaborar índices propios, según las características que por su relevancia se quieran considerar y las políticas que se hayan trazado. No existe una receta universal para diseñar e implementar indicadores adecuados, sino que cada país o institución deberá realizar un diseño propio que garantice que los indicadores producidos sean realmente útiles para la toma de decisiones, justificando por tanto la no despreciable inversión de recursos en su producción y mantención.

A la hora de resumir una medida sintética, la práctica generalizada se centra en resumir la información considerada en un conjunto de indicadores determinados, partiendo de umbrales o valores de referencia, lo que permite usar esas disparidades o distancias respecto a los valores reales, como medidas adimensionales, facilitando su agregación.

Según [5] es importante definir las condiciones que a priori han de exigirse un indicador sintético genérico “I”:

- *Postulado I. Existencia y Determinación.* La función matemática que define el índice ha de existir y tener solución.

¹Se refiere a la integración de varias variables en una sola expresión numérica.

- *Postulado II. Monotonía.* El índice ha de responder positivamente a una modificación positiva de los componentes y negativamente a una modificación negativa. Ello obliga en algunos casos a cambiar el signo de los indicadores cuya correlación sea negativa con el objetivo a medir (desarrollo).
- *Postulado III. Unicidad.* El índice ha de ser único para una situación dada.
- *Postulado IV. Invariancia.* El índice ha de ser invariante respecto a un cambio de origen o de escala de los componentes.
- *Postulado V. Homogeneidad.* La función matemática que define el índice

$$I = f(I_1, I_2, \dots, I_m)$$

ha de ser homogénea de grado uno. Este postulado es vital para la cardinalidad del índice:

$$f(c * I_1, c * I_2, \dots, c * I_m) = c * f(I_1, I_2, \dots, I_m)$$

- *Postulado VI. Transitividad.* Si (a), (b) y (c) son tres situaciones distintas del objetivo medible por el índice, e I(a), I(b) e I(c) son los valores del indicador correspondiente a esas tres situaciones, debe verificarse que:

$$[I(a) > I(b) > I(c)] \Rightarrow [I(a) > I(c)]$$

- *Postulado VII. Exhaustividad.* El índice debe ser tal que aproveche al máximo y de forma útil la información suministrada por los indicadores simples.

Existen cuatro pasos fundamentales para la construcción de índices compuestos o agregados: 1) selección de los indicadores, la escala y la meta, 2) normalización, 3) la ponderación, 4) la agregación de las variables. [19]

La utilización de técnicas multicriterio para la ponderación y agregación de información multidimensional ayuda a evitar los problemas de inconmensurabilidad y pérdida de información que aparecen al expresar los indicadores en una escala común. Sin embargo, estos métodos están llenos de subjetividad en la ponderación de aspectos de la sostenibilidad, por lo que pueden dar como resultado diferentes opciones. En relación a esto, es necesario que estén relacionadas las reglas de agregación de los indicadores simples con las reglas que definen las interrelaciones entre el conjunto de indicadores seleccionados, constituyendo un verdadero modelo de la realidad. Así, se propone la selección de indicadores que representen variables o propiedades del sistema completo, es decir, indicadores holísticos. [5]

3 PROCEDIMIENTO PARA LA EVALUACIÓN DEL DESARROLLO SOSTENIBLE EN LAS PROVINCIAS CUBANAS

A fin de contar con un procedimiento metodológico para evaluar el desarrollo sostenible de forma integral en las provincias cubanas, se diseñó un procedimiento propio, que toma como referencia la metodología Biograma 2008 [26], desarrollada por el Instituto Interamericano de Cooperación para la Agricultura (IICA) y otras similares [19].

3.1. Características del procedimiento

El procedimiento propuesto es un instrumento que tiene por objetivo estimar, y a la vez, representar de manera rápida, el grado relativo de desarrollo sostenible en el territorio, con la expectativa de que dicha estimación provea una base para el diseño de políticas e inversiones orientadas a la aplicación de medidas correctivas y preventivas.

Al aplicar el procedimiento se obtiene el Índice de Desarrollo Sostenible Territorial (IDST) y un sistema de indicadores de desarrollo sostenible organizado mediante un marco ordenador temático, que permite visualizar con más detalle el comportamiento de los indicadores que componen el índice. Ambos instrumentos se complementan para intentar evaluar el estado de desarrollo sostenible del territorio analizado en un período determinado.

El procedimiento se sustenta sobre los principios siguientes:

- a) **Pertinencia:** relacionada con la propuesta de un procedimiento para la evaluación de la sostenibilidad, que se adecua a las condiciones existentes y a la necesidad de las autoridades cubanas de darle seguimiento a las cuestiones de la sostenibilidad en los territorios para actuar oportunamente.
- b) **Racionalidad:** de acuerdo con la relación gasto-beneficio que se requiere para su aplicación.
- c) **Consistencia lógica:** en función de la ejecución de sus pasos en la secuencia planteada y la correspondencia con la lógica de la ejecución de este tipo de estudio.

- d) Generalidad: posibilidad de su extensión como instrumento metodológico para la evaluación del desarrollo sostenible en las provincias cubanas.

3.2. Descripción del procedimiento propuesto

El procedimiento metodológico se ha estructurado en cuatro fases principales que se describen más adelante:

Fase I. Diagnóstico del territorio

Fase II. Estructuración del sistema de indicadores:

- Seleccionar dimensiones, áreas claves e indicadores
- Ponderar indicadores, áreas claves y dimensiones

Fase III. Cálculo del Índice de Desarrollo Sostenible Territorial

- Normalizar indicadores
- Calcular índices de las áreas claves
- Calcular Índices de las dimensiones
- Calcular el IDST

Fase IV. Evaluación de los resultados

- Evaluar la información
- Comunicar los resultados

Fase I. Diagnóstico del territorio

En esta etapa se estudian las características del territorio a evaluar y por consiguiente donde se implementarán estrategias, políticas e inversiones para superar las limitantes responsables de los desequilibrios espaciales.

También se deben determinar las principales potencialidades y vulnerabilidades en la esfera económica, social y ambiental.

Para realizar el diagnóstico podemos obtener información de diversas fuentes:

- Información estadística
- Información institucional
- Bibliografía existente (estudios realizados, monografías, prensa, etc.)
- Entrevistas a especialistas y pobladores
- Observación del territorio

El análisis de la situación ambiental del territorio debe incluir los aspectos ambientales más significativos y los principales factores que inciden en el agotamiento y la contaminación de los recursos naturales.

Fase II. Estructuración del sistema de indicadores

Paso 2.1 Seleccionar las dimensiones y áreas claves

El sistema de indicadores que servirá de base para la evaluación del desarrollo sostenible en el territorio y para el cálculo del IDST se estructura en dimensiones. Las dimensiones se definen como aquellos diversos componentes del sistema que se analizarán, en coherencia con el concepto de desarrollo sostenible. Sin embargo, cabe señalar que las dimensiones podrán variar dependiendo del territorio seleccionado y el fenómeno que se desee analizar.

Las tres dimensiones principales del desarrollo sostenible son: ambiental, económica y social, pero el usuario puede considerar otras dimensiones como la política-institucional y la cultural.

En cada dimensión deben definirse las áreas claves o temas de mayor relevancia para la sostenibilidad, basándose en el diagnóstico realizado en la fase anterior.

Paso 2.2 Seleccionar los indicadores

La elección de los indicadores es una materia delicada, pues deben estar apropiadamente sustentados por bases teóricas relacionadas con cada dimensión analizada, por lo que el criterio y la validación de expertos adquieren importancia en este punto. De manera general, los indicadores seleccionados deben estar en

correspondencia con los principios básicos del desarrollo sostenible: equidad, conservación, eficiencia y gobernabilidad.

Debe elaborarse un listado inicial de indicadores que reflejen coherentemente las variables explicativas para cada dimensión y área clave; deben responder a los objetivos de sostenibilidad que se hayan planteado. Luego se seleccionarán aquellos indicadores de los que se dispongan datos, pues en ocasiones los sistemas estadísticos territoriales no disponen de todos los datos que podrían considerarse relevantes y esto afectaría el cálculo del IDST. Posteriormente se somete el conjunto de indicadores seleccionados en cada dimensión a un análisis multifactorial aplicando la técnica de Análisis de Componentes Principales para reducir el conjunto de variables, eliminando las correlaciones.

Los resultados del análisis multifactorial deben someterse al criterio de los expertos para determinar el conjunto final de indicadores.

A la vez que se seleccionan los indicadores, debe definirse si cada uno tiene una relación positiva o negativa con el desarrollo. Es decir, si el aumento del valor del indicador refleja una situación mejor o peor para la dimensión. Si un aumento en el valor del indicador resulta en una mejoría del sistema, se considera que se tiene una relación positiva. Por el contrario, si un aumento en el valor del indicador empeora la situación, se tiene una relación inversa o negativa [26].

Un aspecto muy importante en esta fase es establecer los umbrales de sostenibilidad o valores de referencia (valores mínimos, máximos, o intervalos) sobre la base de los cuales se comparará el comportamiento de los indicadores. Para alcanzar este propósito se tendrán en cuenta algunos de los siguientes aspectos según el indicador en cuestión:

- a) Normas nacionales e internacionales,
- b) Valores extremos obtenidos del análisis estadístico independiente de las series de datos,
- c) Consideraciones empíricas de los expertos.

Posteriormente se procede a confeccionar la ficha de los indicadores la cual va a permitir almacenar información del indicador como:

- *Nombre del Indicador:* Describe de manera sintética y clara el objetivo del indicador.
- *Dimensión:* Nombre de la dimensión a la que está asociado.
- *Área clave:* Nombre del área temática clave a la que está asociado.
- *Descripción:* se especificará una breve descripción del significado del indicador.
- *Unidad de Medida:* Unidades en la que será expresada el indicador.
- *Fuente:* Fuente de la que se obtienen los datos del indicador.
- *Método de Cálculo:* Indica cuáles son las bases teóricas-científicas en las cuales se basan los métodos de medición requeridas. Representación matemática del indicador.
- *Umbrales de sostenibilidad:* Metas establecidas para el indicador.

Paso 2.3 Determinar los pesos de los indicadores en cada área clave, de las áreas claves en la dimensión y de las dimensiones respecto a la sostenibilidad del sistema.

Para la determinación de los pesos se utiliza el método AHP propuesto por Saaty en 1980 y se emplea el software *Expert Choice versión 11.0*. Se propone el siguiente flujo de trabajo:

1. Construcción del modelo estructural del índice. Para este paso se utilizará el software *Expert Choice*.
2. Construir la matriz de alcance local para el modelo: en este paso se identifican las dimensiones, los indicadores seleccionados en cada una y relaciones causales que puedan existir entre ellos, para esto se hace uso de una tabla binaria. Se pregunta si existe relación entre cada par de nodos de la matriz. En caso de existir relación se asigna el valor 1, sino 0.
3. Emitir los juicios para determinar la importancia de los elementos, perspectivas e indicadores: en este paso se emiten los juicios de los expertos tanto para los nodos de un grupo, como para la relación de un nodo de un grupo con respecto a otro nodo de diferente grupo. En todos los casos la inconsistencia tiene que ser menor que el 10% para que se acepte el juicio de los expertos de lo contrario se procede a emitir nuevamente los juicios.
4. Obtención de la súper-matriz ponderada: es el resultado de la multiplicación de los pesos locales obtenidos en el paso anterior por los pesos del grupo. La suma de sus columnas tiene que ser igual a la unidad.
5. En caso de que las sumas de las columnas no den igual a uno se procede a normalizar por columnas.

6. Obtención de la matriz-límite: se obtiene de elevar a potencias de 2 la súper-matriz ponderada hasta que se estabilice (todas las columnas tienen los mismos valores) siendo los valores de las columnas los pesos relativos por grupos, quedando calculados.

Fase III. Cálculo del Índice de Desarrollo Sostenible Territorial

El IDST sintetiza en un valor numérico el progreso o retroceso hacia el objetivo de desarrollo sostenible, a fin de verificar de forma simple y continuada, la efectividad de las políticas públicas.

El cálculo del IDST se realiza aplicando una suma ponderada de los índices de las dimensiones definidas. Se propone que cuando se hayan seleccionado solamente las tres dimensiones principales del desarrollo sostenible (ambiental, económica y social) los pesos sean iguales para que exista equilibrio entre ellas, pero si el usuario decide incluir otras como la política-institucional y/o la cultural queda a su criterio la ponderación de las mismas, la cual debe estar validada por la consideración de expertos.

Paso 3.1 Normalización de los indicadores

Uno de los principales problemas que se afronta en esta fase es la diversidad de unidades de medidas que pueden encontrarse en los indicadores seleccionados, lo que puede dificultar la integración de los mismos en un índice. Por tanto, debe realizarse un proceso de normalización de todos los indicadores.

Para estandarizar los indicadores en una escala de 0 a 1 se utiliza la siguiente función:

- a) Si el indicador i satisface **"a mayor valor mejor"**

$$IP_i = \begin{cases} \frac{x_i}{V_{Xi}^*} ; x_i < V_{Xi}^* \\ 1 ; x_i \geq V_{Xi}^* \end{cases}$$

- b) Si el indicador i satisface **"a menor valor mejor"**

$$IP_i = \begin{cases} \frac{V_{Xi}^*}{x_i} ; x_i > V_{Xi}^* \\ 1 ; x_i \leq V_{Xi}^* \end{cases}$$

- c) Si el valor deseado para el indicador i es **un intervalo** ($V_{Xi}^* \min; V_{Xi}^* \max$)

$$IP_i = \begin{cases} \frac{x_i}{V_{Xi}^*} ; x_i < V_{Xi \min}^* \\ 1 ; V_{Xi \min}^* \leq x_i \leq V_{Xi \max}^* \\ \frac{V_{Xi}^*}{x_i} ; x_i > V_{Xi \max}^* \end{cases}$$

donde:

IP_i : Valor normalizado del indicador i .

X_i : Valor real alcanzado del parámetro seleccionado para el indicador.

V_{Xi}^* : Valor deseable para el indicador i .

La normalización se puede realizar respecto a los valores de referencia seleccionados para cada indicador, considerándola como una estandarización en términos de distancia respecto al valor objetivo de cada indicador. Estos valores de referencia pueden establecerse considerando límites establecidos por normas nacionales e internacionales, si existen, o los valores máximos o mínimos recogidos en las series de datos disponibles, teniendo en cuenta la tendencia del indicador.

Cuando el valor deseado o el valor real del indicador toman el valor cero, deben sustituirse por un valor muy próximo a él para evitar resultados erróneos.

Tras esta normalización, todos los indicadores son adimensionales, tomando valores en el intervalo (0,1). Un valor próximo a cero indica que, en la dimensión recogida por el indicador específico, se está lejos del objetivo marcado. De forma complementaria, un valor del indicador normalizado próximo a 1 refleja que se cumple el objetivo definido para la consecución del desarrollo sostenible.

Paso 3.2 Cálculo de los índices de cada área clave

Se calcula el valor correspondiente a cada área clave con la siguiente expresión:

$$IAC_j = \sum_{i=1}^n W_{ij} IP_i$$

donde:

IAC_j : Índice del área clave j .

W_{ij} : El peso relativo del indicador i en el área clave j .

IP_i : Valor normalizado del indicador i .

n : Cantidad de indicadores asociados al área clave.

Una vez se hayan calculado los índices de todas las áreas claves se calcula el índice de las dimensiones.

Paso 3.2 Cálculo de los índices de cada dimensión

Se calcula del IDS siguiendo con la siguiente expresión:

$$ID_k = \sum_{j=1}^n W_{jk} IAC_j$$

donde:

ID_k : Índice de la dimensión k .

W_{jk} : El peso relativo del área clave j en la dimensión k .

IAC_j : Valor normalizado del área clave j .

n : Cantidad de áreas claves asociadas a la dimensión.

Paso 3.2 Cálculo del IDST

Se calcula el IDST aplicando la suma ponderada de los índices de las dimensiones:

$$IDST = \sum_{k=1}^n W_k ID_k$$

donde:

IDST: Índice de Desarrollo Sostenible Territorial.

W_k : El peso de la dimensión k .

ID_k : Valor del índice de la dimensión k .

n : Cantidad de dimensiones asociadas al sistema de indicadores.

Fase IV. Evaluación de los resultados

Paso 4.1 Evaluar la información de los indicadores y del desarrollo sostenible territorial

En esta etapa se realiza una valoración de los resultados, la cual debe ser coherente con las metas y el alcance propuesto.

Del análisis de IDST y de los índices que lo componen se podrá determinar cuáles son los puntos críticos que más están afectando al territorio para reorientar los esfuerzos de las administraciones hacia los indicadores en peores condiciones.

El potencial de mejora ofrece una idea del comportamiento de los indicadores en relación a la sostenibilidad en función de su importancia relativa. Para identificar el potencial de mejora del indicador i , perteneciente al área clave j y la dimensión k , se propone la fórmula siguiente:

$$\text{Potencial de mejora}_i = W_k * W_{jk} * W_{ijk} * (1 - IP_i)$$

donde:

W_k : El peso relativo de la dimensión k .

W_{jk} : El peso relativo del área clave j de la dimensión k .

W_{ijk} : El peso relativo del indicador i en el área clave j de la dimensión k .

IP_i : Valor normalizado del indicador i .

Esta fórmula toma en cuenta el complemento del indicador definido, o sea el valor numérico que falta para llegar a la meta y los pesos del indicador, del área clave y de la dimensión.

Para establecer una relación entre el valor numérico del IDST y una evaluación cualitativa del desarrollo sostenible territorial según el cumplimiento de las metas establecidas, se propone utilizar la escala de la Tabla 1 tomando como base la que se aplica en la metodología Biograma 2008 presentada por [26].

Tabla 1. Escala de evaluación del desarrollo sostenible. [29]

Rango	Nivel de evaluación	Color
$0.80 \leq IDST \leq 1.00$	Situación óptima del sistema	Verde
$0.60 \leq IDST < 0.80$	Sistema estable del sistema	Azul
$0.40 \leq IDST < 0.60$	Situación inestable del sistema	Amarillo
$0.20 \leq IDST < 0.40$	Situación crítica del sistema	Anaranjado
$IDST < 0.20$	Alta probabilidad de colapso del sistema	Rojo

Paso 4.2 Comunicar los resultados

La interpretación del IDST puede favorecer y establecer un canal de comunicación efectivo con las partes interesadas en el desarrollo sostenible territorial, ya sean las administraciones, las instituciones o la población del territorio.

En esta etapa debe quedar confeccionado un informe con los principales resultados de la aplicación del procedimiento que ayude a las administraciones al mejoramiento continuo de las estrategias en cada una de las dimensiones.

4 APLICACIÓN DEL PROCEDIMIENTO GENERAL PARA LA EVALUACIÓN DEL DESARROLLO SOSTENIBLE EN LA PROVINCIA DE VILLA CLARA

Para comprobar la factibilidad de la aplicación del procedimiento propuesto para la evaluación del desarrollo sostenible territorial y la identificación de puntos críticos y oportunidades de mejoras que puedan ayudar a las administraciones a reorientar los esfuerzos de su gestión, se aplicó el mismo en la provincia Villa Clara.

El procedimiento propuesto está sustentado computacionalmente por una herramienta informática desarrollada también como parte de esta investigación, que fue empleada en la aplicación práctica del procedimiento. En [29] se pueden obtener los detalles de las características y el diseño del software. Esta herramienta constituye un valioso instrumento para apoyar el almacenamiento, recuperación e integración de los indicadores de desarrollo sostenible, facilita el cálculo y la representación gráfica del indicador global y del sistema de indicadores en su totalidad, y resuelve una de las carencias actuales en la evaluación de la sostenibilidad: el soporte sobre herramientas informáticas.

A continuación, se detalla la aplicación del procedimiento a la provincia de Villa Clara.

Fase I. Diagnóstico del territorio

La provincia de Villa Clara está situada en la región central de Cuba, entre los 22°16' y 23°09' de latitud norte y los 80°02' y 80°25' de longitud oeste. Limita al norte con el Océano Atlántico, al este con la provincia Sancti Spíritus, al sur con la provincia Sancti Spíritus y al oeste con las provincias Matanzas y Cienfuegos. Posee una extensión territorial de 8411,81 km², incluidos 492,4 km², de cayos adyacentes lo que la ubica en el quinto lugar por extensión entre las 15 provincias del territorio nacional, su extensión representa el 7,6 % del área total del país. Las costas, que poseen características bajas y pantanosas, alcanzan una longitud de 191,0 kilómetros por el norte y son la única frontera marítima. El relieve se caracteriza por las alturas del Norte de Cuba Central, llanura de Manacas y las alturas de Santa Clara. Su hidrografía está representada por los ríos Sagua la Grande y Sagua la Chica y el embalse Alacranes, Hanabanilla, Minerva y Palma Sola. [21]

Según datos del anuario estadístico provincial del año 2014 Villa Clara posee una población total de 792 292 habitantes, para una densidad de población 94.2 habitantes por km² y es una de las provincias más envejecidas del país.

Los problemas ambientales de mayor incidencia e impacto en el territorio, que deben enfrentarse en el camino hacia el desarrollo sostenible son [9]:

- Degradación de los suelos
- Afectaciones a la cobertura forestal
- Contaminación (residuales líquidos, sólidos, emisiones a la atmósfera, productos químicos y desechos peligrosos)
- Pérdida de diversidad biológica
- Dificultades con la disponibilidad y calidad del agua
- Impactos del cambio climático

Estos problemas se manifiestan de forma negativa en la consecución de la elevación sostenida y equitativa de la calidad de vida de las personas, por sus impactos económicos y el efecto sobre los ecosistemas y los recursos naturales.

Las principales actividades económicas del territorio son la agricultura y el turismo, aunque también se desarrolla la industria y otras. Se destacan las producciones de café, tabaco, azúcar y cultivos varios. El aeropuerto internacional “Abel Santamaría Cuadrado” de la capital provincial recibe más de 80 vuelos semanales en temporada alta y más de 60 en temporada baja provenientes de varios países de Europa y América, con turistas que se dirigen al polo turístico de la cayería norte villaclareña.

Fase II. Estructuración del sistema de indicadores

El sistema de indicadores que servirá de base para la evaluación del desarrollo sostenible en el territorio villaclareño y el cálculo del IDST, deberá garantizar que se mida el cumplimiento de los principios básicos de: equidad intrageneracional e intergeneracional, conservación, y eficiencia.

El sistema se estructura de acuerdo a las tres dimensiones principales del desarrollo sostenible: ambiental, económica y social.

Se definen como áreas claves de la dimensión ambiental, tomando en consideración los principales problemas ambientales y las prioridades para la sostenibilidad del territorio las siguientes:

- Protección y rehabilitación de los suelos
- Contaminación atmosférica
- Pérdida de la diversidad biológica
- Gestión de recursos hídricos

En el orden social se decidió incluir en la evaluación las siguientes áreas claves por la connotación que tienen para la sostenibilidad de toda sociedad:

- Salud
- Educación
- Equidad

En el ámbito económico se seleccionaron las siguientes áreas claves:

- Desarrollo económico
- Uso de la energía
- Generación de desechos
- Transporte

Una vez identificados los aspectos claves de cada dimensión se procedió a identificar un listado inicial de indicadores que podrían formar parte del sistema.

Este listado se refinó en primera instancia seleccionando aquellos indicadores de los que existen series temporales de datos oficiales publicados por la Oficina Nacional de Estadística e Información (ONEI) o datos ofrecidos por otras instituciones. Posteriormente se aplicaron técnicas de análisis multifactorial para identificar las posibles correlaciones entre indicadores y eliminar redundancias. Una vez estructurado el sistema de indicadores se procedió a establecer los valores de referencia y los pesos relativos de los indicadores, áreas claves y dimensiones en el sistema. Para el cálculo de los pesos se utilizó el método AHP y el software *Expert Choice versión 11.0*. Se obtuvo la siguiente jerarquía y los pesos relativos de las dimensiones, áreas claves e indicadores (ver Figura 1).



Figura 1. Jerarquía del sistema de indicadores y pesos relativos [29]

Fase III. Cálculo del IDST

Con los datos disponibles se procede a crear el sistema de indicadores en la herramienta informática desarrollada en este trabajo, para que sirva de base en las próximas fases.

En etapa se calcula el IDST de la provincia Villa Clara para el período 2008-2012, con el objetivo de analizar el grado de sostenibilidad relativa alcanzado en el territorio en cinco años consecutivos (ver Figura 2). En la Tabla 2 se detallan los índices por dimensiones y áreas claves para el período seleccionado.

Sistema para la Evaluación del Desarrollo Sostenible

Inicio | Nomencladores | Sistemas de indicadores | Captura de datos | Reportes | Usuarios | Cerrar sesión

Inicio » Realizar comparación Espacio-Temporal

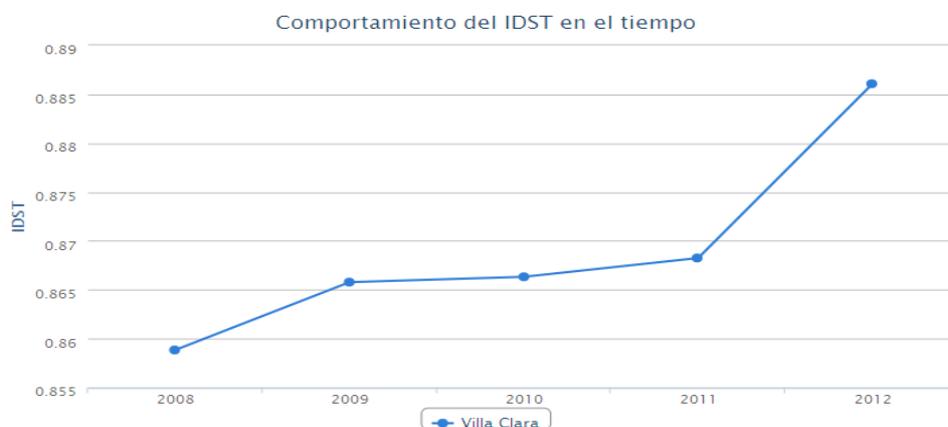


Figura 2. Comportamiento del IDST 2008-2012 en Villa Clara [29]

Tabla 2. Índices por dimensiones y áreas claves para Villa Clara 2008-2012 [29]

Territorio	Villa Clara	2008	2009	2010	2011	2012
Dimensiones	Áreas Claves					
Ambiental		0.86	0.86	0.87	0.87	0.88
	Protección y rehabilitación de los suelos	0.92	0.93	0.94	0.94	0.95
	Pérdida de la diversidad Biológica	0.66	0.66	0.66	0.66	0.69
	Gestión de recursos hídricos	1.00	1.00	1.00	1.00	1.00
Económica		0.74	0.77	0.76	0.76	0.81
	Desarrollo económico	0.80	0.82	0.80	0.86	0.83
	Uso de la energía	0.72	0.86	0.86	0.79	0.85
	Transporte	0.47	0.43	0.43	0.40	0.60
Social	Generación de desechos	1.00	1.00	1.00	1.00	1.00
		0.97	0.96	0.97	0.98	0.97
	Equidad	0.99	0.97	0.99	0.99	0.98
	Salud	0.98	0.97	0.97	0.98	0.96
IDST	Educación	0.95	0.94	0.96	0.96	0.96
		0.86	0.87	0.87	0.87	0.89
Evaluación		Situación óptima del sistema				
Leyenda		Situación Óptima	Situación Estable	Situación Inestable	Situación Crítica	Afuera del alcance

En las siguientes ilustraciones se muestran los gráficos que genera el software para el territorio villaclareño en el año 2012. La Figura 3 muestra el Biograma, los índices de las dimensiones se observan en la Figura 4, y la

Tabla 3 presenta el sistema de indicadores totalmente desglosado.

Sistema para la Evaluación del Desarrollo Sostenible

Inicio | Nomencladores | Sistemas de indicadores | Captura de datos | Reportes | Usuarios | Cerrar sesión (admin)

Inicio » Calcular IDST » Villa Clara 2012

Villa Clara 2012

IDST = 0.89 Evaluación: Situación óptima del sistema

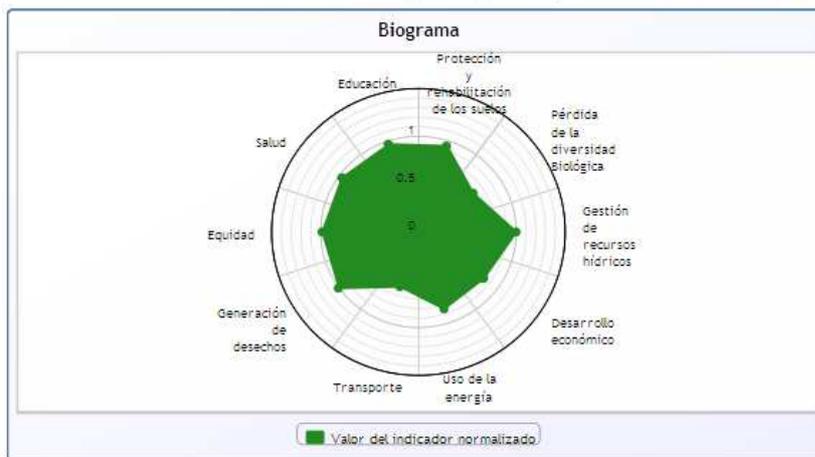


Figura 3. Biograma obtenido para la provincia Villa Clara en el año 2012 [29]

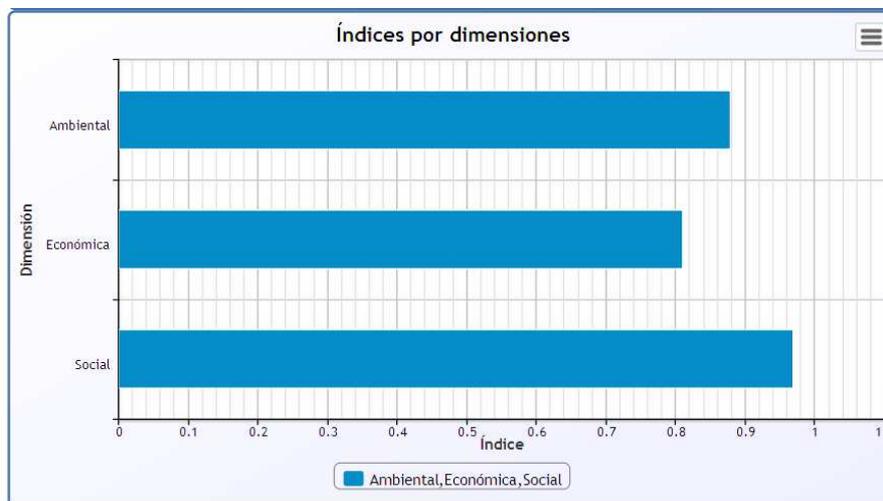


Figura 4. Índices por dimensiones: Villa Clara 2012 [29]

Tabla 3. Sistema de indicadores Villa Clara 2012 [29]

Dimensión:	Económica	Índice de la dimensión:	0.81
Área Clave:	Desarrollo económico	Índice del Área Clave:	0.83
Indicador	Valor	Valor normalizado	
Tasa de crecimiento Producción Mercantil	98.8 %		0.99
Volumen de inversiones per cápita	160.14 MP/habitante		1.00
Inversiones Ciencia y Técnica	0.0309 %		0.15
Área Clave:	Uso de la energía	Índice del Área Clave:	0.85
Indicador	Valor	Valor normalizado	
Intensidad energética	0.072 tcc/MP		0.94
Energía sustit fuentes alternativas	864.4 tep		0.60
Consumo de biomasa en relación a la producción	99.9 %		1.00
Área Clave:	Transporte	Índice del Área Clave:	0.60
Indicador	Valor	Valor normalizado	
Pasajeros transp. ómnibus públicos/hab.	0.0366 Mpasajeros/habitante		0.20
Tráfico de carga	314.8 Millones de toneladas-km		1.00
Área Clave:	Generación de desechos	Índice del Área Clave:	1.00
Indicador	Valor	Valor normalizado	
Volumen de desechos recogidos/ hab/día	1.71 m3/habitante		1.00

Inicio » Sistema de Indicadores » Villa Clara 2012

Sistema de indicadores	Villa Clara 2012	IDST	0.89
Dimensión:	Ambiental	Índice de la dimensión:	0.88
Área Clave:	Protección y rehabilitación de los suelos	Índice del Área Clave:	0.85
Indicador	Valor	Valor normalizado	
Índice de boscosidad	22.8 %		0.86
Áreas beneficiadas por el PNMCS/plan	102 %		1.00
Áreas beneficiadas PNMCS Cuencas/plan	107 %		1.00
Áreas beneficiadas PNMCS Plan Turq/plan	103 %		1.00
Área Clave:	Pérdida de la diversidad Biológica	Índice del Área Clave:	0.69
Indicador	Valor	Valor normalizado	
Áreas terrestres protegidas/ST	5.15 %		0.14
Áreas marinas protegidas/ST	22.08 %		1.00
Dinámica de la captura bruta total de peces plataforma	107.6 %		0.93
Área Clave:	Gestión de recursos hídricos	Índice del Área Clave:	1.00
Indicador	Valor	Valor normalizado	
Carga contaminante dispuesta /hab	0.011 T DBO/año		1.00
Extracción agua dulce con relación a la disponible	79.53 %		1.00

Dimensión:	Social	Índice de la dimensión:	0.97
Área Clave:	Equidad	Índice del Área Clave:	0.98
Indicador	Valor	Valor normalizado	
Tasa de desocupación	3.4 %		1.00
Tasa desoc mujer/hombres	1.03 adimensional		0.97
Participación en las últimas elecciones APPP	91.9 %		0.92
Dotación normal de camas en unidades de asistencia social	1381 u		1.00
Área Clave:	Salud	Índice del Área Clave:	0.96
Indicador	Valor	Valor normalizado	
Tasa de mortalidad infantil < 5 años	6.1 /1000 nacidos vivos		1.00
Cobertura de saneamiento	100 %		1.00
Cobertura de abasto de agua potable	100 %		1.00
Índice de bajo peso al nacer	4.2 /100 nacidos vivos		0.82
Esperanza de vida al nacer	79.43 años		1.00
Área Clave:	Educación	Índice del Área Clave:	0.99
Indicador	Valor	Valor normalizado	
PEA con nivel educacional MS o S	67.58 %		0.94
Retención en el ciclo normal de la enseñanza primaria	98.6 %		0.99

Leyenda Situación Óptima Situación Estable Situación Inestable Situación Crítica Al borde del colapso

En la Figura 5 se muestran los potenciales de mejora para Villa Clara en el año 2012.

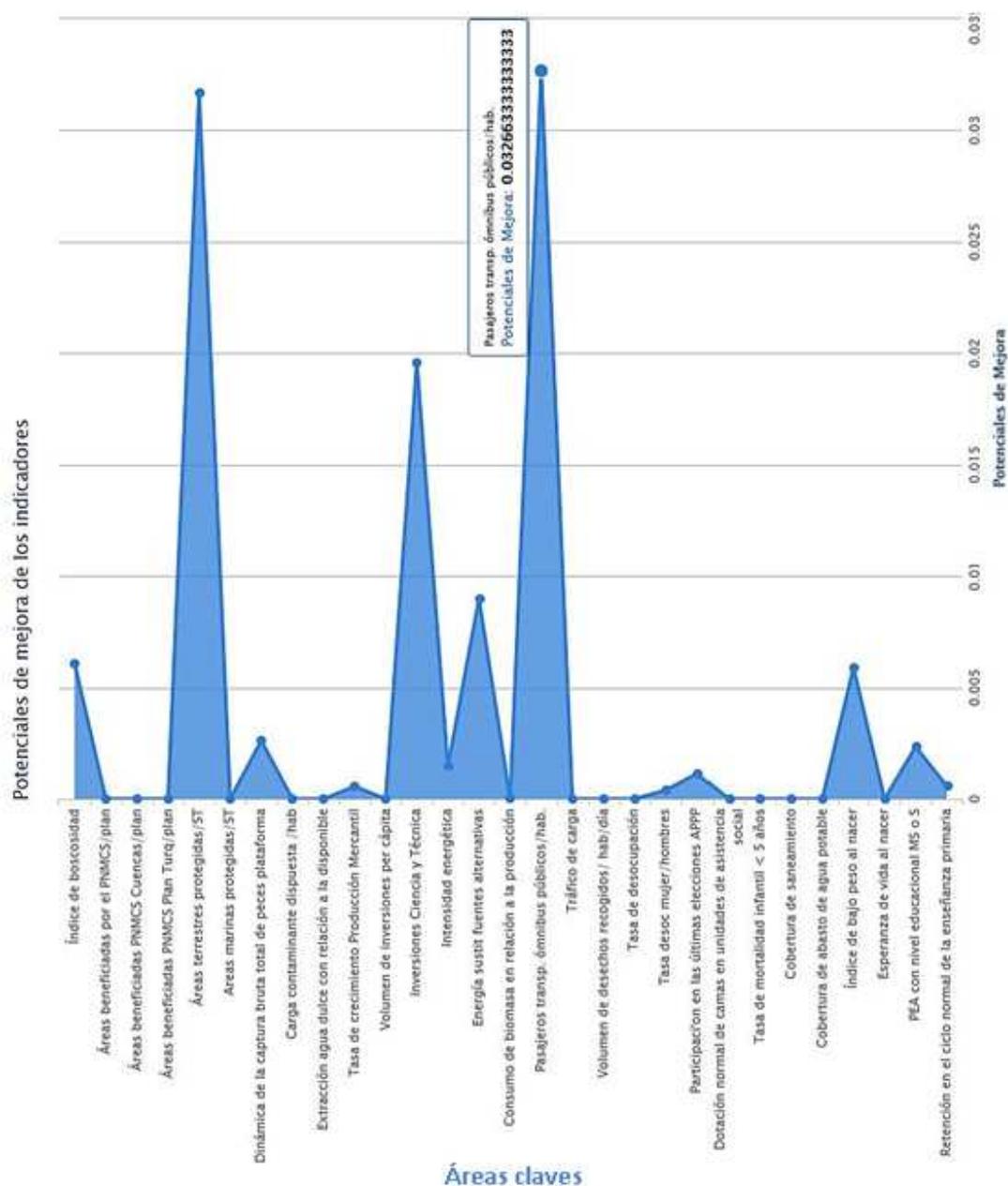


Figura 5. Potenciales de mejora. Villa Clara 2012 [29]

Fase IV. Evaluación de los resultados

Como se observa en la Figura 2, el IDST tuvo una evolución favorable en el período 2008-2012 aumentando su valor de 0.86 a 0.89. De acuerdo al sistema de indicadores diseñado y las metas de sostenibilidad definidas el territorio cumple los parámetros de una situación óptima para la sostenibilidad, la cual mantiene se mantiene durante los cinco años estudiados. De manera general, se observan avances en la esfera ambiental, económica y social.

Al analizar los potenciales de mejora se aprecia que las mayores dificultades se observan en el orden económico, específicamente en el área clave del Transporte por el valor que alcanza el indicador Pasajeros transportados en ómnibus de transporte público por habitante y en el área de desarrollo económico se ve

afectado el Volumen de inversiones en la esfera de la ciencia y la innovación tecnológica que es garantía para el desarrollo en esta rama.

En la esfera ambiental, es necesario potenciar las áreas terrestres protegidas pues a pesar de que ha ido incrementándose el valor de la proporción respecto a la superficie total del territorio aún está un poco alejado del alcanzado en otras provincias.

La dimensión social es la de mejores resultados y los potenciales de mejora de sus indicadores son bastante bajos.

5 CONCLUSIONES

El procedimiento propuesto establece una línea de acción para seleccionar, recopilar, analizar, integrar y evaluar los indicadores de desarrollo sostenible territorial. Permite que la evaluación de sostenibilidad sea complementada con medidas de resultados para que las administraciones territoriales puedan conocer si las políticas, estrategias y metas ambientales, económicas y sociales son efectivas. En él se evidencian las siguientes características:

- Pertinencia: Se adecúa a las condiciones existentes y satisface las necesidades de las administraciones territoriales de evaluar la sostenibilidad del desarrollo en el territorio para mejorar su gestión y apoyar la toma de decisiones. Permite identificar los puntos críticos y oportunidades de mejora en el territorio objeto de estudio sobre la base de un indicador global de desarrollo sostenible y los índices que lo componen.
- Consistencia lógica: La secuencia lógica de las etapas dentro de cada fase del procedimiento, es mantenida independientemente del territorio donde se aplique.
- Generalidad: El procedimiento propuesto, se considera lo suficientemente flexible y adaptable para ser generalizable a otros territorios que deseen evaluar su grado relativo de desarrollo sostenible.

La aplicación del procedimiento en la provincia objeto de estudio permitió revelar los principales problemas e identificar las oportunidades de mejora del desarrollo sostenible, demostrando así su validez y efectividad. Por tanto, se propone generalizar su uso al resto de los territorios del país con el fin de que se pueda arribar a conclusiones sobre el estado en que se encuentran en relación al desarrollo sostenible en cada uno de ellos, así como realizar comparaciones entre provincias de similares condiciones.

Resulta interesante continuar perfeccionando la aplicación informática desarrollada haciendo uso de otras posibilidades de procesamiento de información, como podría ser la introducción de técnicas de análisis multivariante para el estudio de los resultados.

REFERENCIAS

- [1] Achkar, Marcel. (2005). Ordenamiento Ambiental del Territorio. Montevideo: Comisión Sectorial de Educación Permanente.
- [2] AMA, CITMA, & PNUMA. (2009). GEO Cuba: Evaluación del medio ambiente cubano. La Habana.
- [3] Arnés Prieto, Esperanza. (2011). Desarrollo de la metodología de evaluación de sostenibilidad de los campesinos de montaña en San José de Cusmapa (Nicaragua). (Master), Universidad Politécnica de Madrid.
- [4] Bossel, H. (1999). Indicators for Sustainable Development: Theory, Method, Application. Winnipeg, Canada: International Institute for Sustainable Development.
- [5] Castro Bonaño, J. Marcos. (2002). Indicadores de Desarrollo Sostenible Urbano. Una aplicación para Andalucía. (Tesis Doctoral), Universidad de Málaga, Málaga. España. Disponible en <http://www.eumed.net/tesis/>
- [6] CDS. (1995). Indicators of sustainable development: guidelines and methodologies.
- [7] CITMA, Ministerio de Ciencia, Tecnología y Medio Ambiente., & AMA, Agencia de Medio Ambiente. (2005). Situación Ambiental Cubana 2004. Paper presented at the V Convención Internacional sobre Medio Ambiente y Desarrollo, Palacio de las Convenciones de La Habana.
- [8] CITMA, ONE, & PNUMA. (2009). Iniciativa Latinoamericana y Caribeña para el Desarrollo Sostenible: Indicadores de seguimiento: Cuba 2009. (1ª ed. ed.). Ciudad de la Habana.
- [9] CITMA, Delegación Provincial Villa Clara. (2011). Estrategia Territorial de Medio Ambiente Provincia Villa Clara 2011-2015.

- [10] CITMA, MINREX, MINCEX, ACNU. (2012). Informe de Cuba a la Conferencia de Naciones Unidas sobre Desarrollo Sostenible Río +20. La Habana.
- [11] Durán Romero, Gemma. (2000). Medir la sostenibilidad: indicadores económicos, ecológicos y sociales. Madrid: Universidad Autónoma de Madrid.
- [12] Fernández Latorre, Francisco. (2006). Indicadores de sostenibilidad y medio ambiente: métodos y escalas. (C. d. M. A. J. d. Andalucía Ed.). Sevilla.
- [13] Gallopín, G. C. (2003). Sostenibilidad y desarrollo sostenible: un enfoque sistémico (N. Unidad Ed.). Santiago de Chile.
- [14] Leiva Más, Jorge, Rodríguez Pico, Iván L., & Martínez Nodal, Pastora. (2012). Indicadores de sostenibilidad ambiental en universidades: una herramienta para la identificación y evaluación de soluciones tecnológicas. *Afinidad* LXIX, 558(Abril-Junio 2012).
- [15] Márquez, Luiz, & Cuétara, Leonardo (2008). Propuesta y aplicación de un sistema de indicadores para determinar el Índice de Desarrollo Sostenible Global (IDSG) de un destino turístico. Caso: Patanemo, Venezuela. *FACES XVII*(N° 1).
- [16] Martínez de la Vega, Vivian, Pérez Dorta, Frank, Alonso Guerra, Ángel, & Duque Jova, María Cristina. (2014). Diseño e implementación de sistema de indicadores de sostenibilidad turística para instalaciones y destinos de Jardines del Rey. Paper presented at the Forum de Ciencia y Técnica 2014, Escuela de Hotelería y Turismo "Alberto Delgado Delgado".
- [17] Martínez Vega, Javier , Echavarría Daspert, Pilar, González Cascón, Victoria, & Martínez Cruz, Natalia. (2009). Propuesta metodológica para el análisis de la sostenibilidad en la provincia de Cuenca. *Boletín de la A.G.E.*, N.º 49, 281-308.
- [18] Meadows, D. (1998). *Indicators and Information Systems for Sustainable Development: The Sustainability Institute*.
- [19] Medel González, Frank. (2012). Procedimiento para la evaluación del desempeño ambiental. Aplicación en centrales eléctricas del la UEB de generación distribuida de Villa Clara. . (Master), Universidad Central "Martha Abreu" de Las Villas, Santa Clara.
- [20] ONEI. (2009). Medio Ambiente. Estadísticas en la Revolución.
- [21] ONEI. (2014). Anuario Estadístico de Villa Clara 2013.
- [22] ONEI. (2009). Cuba en cifras. Objetivos de desarrollo del milenio. 1990-2008. La Habana.
- [23] Pérez Albert, Yolanda, & Nel Lo Endreu, Marta. (2013). Propuesta de indicadores para evaluar la sostenibilidad de la actividad turística. El caso del Valle de Viñales (Cuba). *Anales de Geografía*, 33(núm. 1 193-210).
- [24] Quiroga Martínez, Rayén. (2007). Indicadores ambientales y de desarrollo sostenible: avances y perspectivas para América Latina y el Caribe. Santiago de Chile: Naciones Unidas. CEPAL.
- [25] Salina Chávez, Eduardo, & La O Osorio, José Alberto. (2006). Turismo y sustentabilidad: de la teoría a la práctica en Cuba. *Cuadernos de Turismo*, enero-junio(017), 201-221.
- [26] Sepúlveda S., Sergio (2008). *Biograma: metodología para estimar el nivel de desarrollo sostenible de territorios*. San José, Costa Rica: IICA.
- [27] United Nations, Division for Sustainable Development. (2001). *Indicators of Sustainable Development: guidelines and methodologies*. New York: United Nations.Division for Sustainable Development.
- [28] United Nations, Sustainable Development. (1992). AGENDA 21. Paper presented at the United Nations Conference on Environment & Development, Rio de Janeiro, Brazil.
- [29] Vega Calcines, Anabel. (2014). Procedimiento para la evaluación del desarrollo sostenible territorial. Aplicación en la provincia de Villa Clara. Universidad Central "Martha Abreu" de Las Villas, Santa Clara.

Capítulo 13

pp 160-165

EVOLUCIÓN DE LAS CONDICIONES SOCIOECONÓMICAS DE HOGARES BENEFICIADOS CON EL PROGRAMA PENSIÓN PARA ADULTOS MAYORES (PPAM) EN TABASCO, MÉXICO. PERÍODO 2013-2015

José Félix García Rodríguez*, Carlos M. Lezama León**, Carlos N. Bouza Herrera***, Ignacio Caamal Cahuich**** Rilton G. Bonfin Primo*****

*Universidad Juárez Autónoma de Tabasco, México. DACEA

**Universidad Juárez Autónoma de Tabasco, México. DACEA

***Universidad de la Habana, Cuba. Facultad de Matemática

**** Universidad Autónoma de Chapingo. Facultad de Economía

*****Centro de Estudios por la Amistad de Latinoamérica, Asia y Brasil (CEALA), Salvador Bahía, Brasil.

ABSTRACT

Aging is a global phenomenon that over time becomes more relevant in all countries for their economic and social implications. In Mexico, the National Council of Population (CONAPO) reported that the group of older adults (65 and over) will increase from 7.1 million in 2010 to 9.8 in 2020, and 23.1 million in 2050 (<http://www.conapo.gob.mx>). Thus, in the not too distant future Mexico is a country of older adults, but more worrying is the tendency to poverty and vulnerability of this population segment. And the economic and social landscape of a poor country with a majority of vulnerable older adults in sight complicated, both for society in general and the state.

KEYWORDS: Poverty, multidimensional poverty, vulnerability, elderly

RESUMEN

El envejecimiento es un fenómeno mundial que con el paso del tiempo toma mayor relevancia en todos los países por sus implicaciones económicas y sociales. En México, el Consejo Nacional de la Población (CONAPO, 2015) informó que el grupo de adultos mayores (65 años y más) pasará de 7.1 millones en 2010 a 9.8 millones en 2020, y a 23.1 millones en 2050, (<http://www.conapo.gob.mx>). Así, en un futuro no muy lejano México será un país de adultos mayores, pero más preocupante es la tendencia a la pobreza y vulnerabilidad de este segmento de población. Y es que el panorama económico y social de un país pobre y con una mayoría de adultos mayores vulnerables se vislumbra complicado, tanto para la sociedad en general como para el Estado.

1. INTRODUCCIÓN

El fenómeno del envejecimiento población cobra cada vez mayor relevancia a nivel internacional y México no es la excepción. En nuestro país, los adultos mayores de hoy día nacieron en una época en la que la mayor parte de la población vivía en condiciones de pobreza. De acuerdo con Székely (2005), en 1950 el 88% de la población se encontraba en pobreza patrimonial, 73% en pobreza de capacidades y 61% en pobreza alimentaria. Además, enfrentaron insuficiencias en provisión de servicios educativos y de salud. Ello dio como resultado varias generaciones con un alto grado de analfabetismo, bajos niveles de salud y alta incidencia de desnutrición.

De esta manera, el envejecimiento tiene múltiples facetas que exigen su análisis y demandan su atención. Normalmente los adultos mayores enfrentan entre otras privaciones problemas de salud, educación, desempleo, baja cobertura en pensiones y protección social, lo que conlleva a problemas de vulnerabilidad, pobreza y desigualdad.

1.1 Vulnerabilidad, pobreza y desigualdad

Es importante conocer las diferencias y similitudes entre vulnerabilidad, pobreza y desigualdad. El Informe de Desarrollo Mundial del Banco Mundial (2001), asocia la vulnerabilidad con la probabilidad de estar expuesto a una gran variedad de riesgos (violencia, crimen, desastres naturales, abandono escolar, etc.). Diversos grupos sociales son vulnerables por que poseen una desventaja por limitaciones físicas, de edad, genero, etnia o pobreza. Más allá de las condiciones de pobreza, pobreza extrema y marginación, existen un sinnúmero de situaciones que comprometen severamente el desarrollo físico e intelectual de las personas. El abandono, la orfandad, la desprotección o desamparo, la indigencia y los diferentes tipos de discapacidades, etc.

En este sentido, el concepto de vulnerabilidad refiere a aquella diversidad de situaciones que pueden sucederle a un individuo o grupo y al proceso por el cual se está en riesgo. Vulnerabilidad no es exactamente lo mismo que pobreza sin embargo esta última la incluye. La pobreza hace referencia a una situación de carencia efectiva y actual, mientras que la vulnerabilidad trasciende esta condición proyectando a futuro la posibilidad de padecerla a partir de ciertas debilidades que se constatan en el presente.

Hemos mencionado anteriormente que no todos los grupos vulnerables son pobres pero si todos los pobres son vulnerables. De acuerdo a la metodología oficial de medición de pobreza multidimensional aplicada por el Consejo Nacional de Evaluación de la Política Social (CONEVAL, 2015), pobres son las personas y los hogares cuyos ingresos son insuficientes para adquirir los bienes y servicios que se requieren para satisfacer sus necesidades, y que presentan carencias en al menos uno de los siguientes seis indicadores: rezago educativo, acceso a los servicios de salud, acceso a la seguridad social, calidad y espacios de la vivienda, servicios básicos en la vivienda y acceso a la alimentación.

Asimismo, se considera pobre extremo cuando tiene tres o más carencias de seis posibles y que además se encuentra por debajo de la línea de bienestar mínimo. Las personas en esta situación disponen de un ingreso tan bajo que aún si lo dedicasen por completo a la adquisición de alimentos, no podrían adquirir los nutrientes necesarios para tener una vida sana, por lo que eliminar la pobreza extrema constituye una de las grandes deudas que se tiene con la sociedad, por lo que es necesario desarrollar esfuerzos conjuntos entre el Estado y la sociedad en general.

Al igual que la vulnerabilidad, la desigualdad está presente no solo entre los pobres, sin embargo todos los pobres viven en desigualdad. No debe confundirse la pobreza con la desigualdad, la pobreza se refiere a la situación económica, la desigualdad es social y política. La pobreza como concepto hace referencia a la carencia de una persona o población por no poder cubrir sus necesidades básicas. En cambio la desigualdad es la imposibilidad de acceder a determinados recursos, derechos, obligaciones, beneficios etc. De esta manera, la desigualdad es una manifestación de la pobreza. La desigualdad puede ser social y económica, de género, racial, cultural y geográfica. La brecha entre los más ricos y los más pobres de un determinado lugar permite establecer el nivel de desigualdad de un país o sociedad, y esto acentúa los niveles de pobreza.

Piketty (2014) plantea que la distribución de la riqueza constituye en la actualidad uno de los temas más debatidos y controversiales dentro de la economía política contemporánea, y que los debates acerca del tema están más cargados de especulaciones teóricas y prejuicios políticos y sociales que de información y hechos objetivos.

La pobreza y la desigualdad social son dos conceptos muy relacionados, para algunos expertos el problema no es la pobreza sino la desigualdad social. Otros consideran que ambos deben ser tratados para lograr tener una sociedad justa y equilibrada, ya que no alcanza con el crecimiento económico, sino que además hay que implementar acciones de distribución de la riqueza e inversión social para lograr un desarrollo humano extendido. Al respecto, (García, 2015, p. 31) menciona que “es imprescindible que haya crecimiento económico, estabilidad, competitividad, eficiencia económica junto con desarrollo social; no son aspectos antitéticos, al contrario. La experiencia de los países exitosos indica que el desarrollo social es un poderoso motor del crecimiento económico”.

En México, el bajo nivel de ingreso y las bajas tasas de ahorro han reducido a la mayor parte de los adultos mayores a una situación de pobreza, condición que se agravará en los próximos años debido a la transformación de la pirámide poblacional que implica una mayor cantidad de adultos mayores respecto a la población total. En consecuencia, resulta de suma importancia elaborar políticas públicas adecuadas que atiendan de manera efectiva a este grupo de población tanto en zonas urbanas como rurales.

La condición de vulnerabilidad de los adultos mayores no solo afecta a este grupo etario, sino también a los hogares en los que residen. El hecho de que los hogares en condiciones de pobreza no cuenten

con los recursos necesarios para enfrentar gastos en situaciones catastróficas en salud, da lugar a que tengan que deshacerse de sus bienes y patrimonio, teniendo una disminución irrecuperable de sus activos ante gastos imprevistos, aumentando así la posibilidad de caer en condiciones de pobreza o de perpetuar la transmisión intergeneracional de la misma.

En este contexto, y con el propósito de mejorar la calidad de vida de los adultos mayores, En México el Gobierno Federal a través de la Secretaría de Desarrollo Social (SEDESOL), puso en marcha el Programa Pensión para Adultos Mayores (PPAM), el cual atiende a las personas adultas mayores de 65 años en adelante, con cobertura nacional. Los beneficiarios del mismo reciben apoyos económicos de 580 pesos mexicanos mensuales pagaderos de forma bimestral. El PPAM contribuye a disminuir las condiciones de pobreza y vulnerabilidad de los Adultos Mayores, tomando en cuenta la vulnerabilidad por seguridad social y la vulnerabilidad por ingresos.

Esta nueva estrategia para la seguridad social amplió la cobertura del programa al aumentar el rango de 70 años a una edad mínima de 65 años para los adultos mayores que no reciban ingresos por concepto de pensión o jubilación de tipo contributivo. Este programa busca asegurar un ingreso mínimo y apoyos de protección social a las personas adultas mayores de 65 años de edad que no reciben un ingreso mensual superior a \$1,092 pesos por concepto de jubilación o pensión de tipo contributivo. El PPAM ha demostrado que la combinación de la entrega de una pensión monetaria de tipo no contributivo, más la realización de acciones para aminorar el deterioro de la salud física y mental a través de la constitución y fortalecimiento de una Red Social en favor de este grupo social, aunados a acciones de protección social coordinadas entre todo el sector público, mejoran efectivamente las condiciones de vida de las personas Adultas Mayores.

2. ASPECTOS METODOLÓGICOS DE LA INVESTIGACIÓN

Esta investigación tiene como propósito llevar a cabo un análisis comparativo de la situación de pobreza y vulnerabilidad de los hogares con adultos mayores del municipio de Centro, Tabasco, México durante el período 2013-2015, buscando confirmar la hipótesis de que dichos hogares han mejorado tales condiciones durante dicho periodo gracias a su incorporación al Programa Pensión para Adultos Mayores.

Particularmente, se busca dar respuesta a las siguientes preguntas de investigación:

- a) ¿Cuál es el bienestar económico de los hogares beneficiarios del Programa Pensión para Adultos Mayores?
- b) ¿Qué carencias sociales se han logrado disminuir en los hogares investigados a partir de la aplicación del Programa?
- c) ¿Cuántos de los hogares beneficiarios del Programa aún se encuentran en pobreza extrema?

Asimismo, el objetivo general de la investigación es conocer la situación socioeconómica de los hogares beneficiarios del Programa Pensión para Adultos Mayores del municipio de Centro Tabasco, con el propósito de retroalimentar la aplicación del Programa en beneficio de este segmento de población altamente vulnerable.

En congruencia con lo anterior, la hipótesis de investigación es: Los hogares beneficiarios del Programa Pensión para Adultos Mayores en el municipio de Centro, Tabasco que se encontraban en pobreza extrema en el año 2013, lograron mejorar sus condiciones socioeconómicas en el año 2015.

2.1 Enfoque, tipo y diseño de la Investigación

Se llevó a cabo una investigación con enfoque cuantitativo y de tipo descriptivo no experimental, con un estimador pre-post a partir de un universo poblacional y la aplicación de un muestreo aleatorio simple. Al respecto, Hernández, Fernández y Baptista (2006, p. 5) expresan que “el enfoque cuantitativo usa la recolección de datos para probar hipótesis, con base en la medición numérica y el análisis estadístico, para establecer patrones de comportamiento y probar teorías”.

La población objeto de estudio se definió como los hogares con adultos mayores en pobreza extrema del municipio de Centro, Tabasco, que fueron incorporados en el año 2013 al PPAM y que al momento de inscribirse al programa tuvieran una edad de entre 65 y 70 años. De esta manera, se determinó un universo poblacional de 4,122 beneficiarios. Asimismo, se determinó como marco muestral a 314 hogares que fueron identificados en pobreza extrema.

El tamaño adecuado de la muestra representativa de la población objetivo de estudio es de 173 hogares con adultos mayores y se determinó a partir de los siguientes criterios: 1) Personas en condición de

pobreza extrema; 2) nivel deseado de confianza, y 3) margen de error aceptable. Dicho tamaño de muestra se calculó a través de la aplicación de la siguiente fórmula:

$$n = \frac{N z^2 p q}{e^2 (N-1) + z^2 p q}$$

Donde:

n= Tamaño de la muestra.

N= Tamaño de la población.

Z= Nivel de confianza del 95%; [1.96]

p= Probabilidad de éxito es de un 50% (0.50)

q= Probabilidad de fracaso 100% - 50% = (0.50)

e= Precisión (Error máximo admisible en términos de proporción) 5%

Como instrumento de recolección de la información de campo se aplicó un cuestionario adaptado a las necesidades de la investigación, a partir del “Cuestionario Único de Información Socioeconómica (CUIS)”, instrumento usado para la evaluación de criterios de identificación de beneficiarios de los programas sociales que tiene a su cargo la Secretaría de Desarrollo Social (SEDESOL). Dicho instrumento es aplicado para dar cumplimiento a los “Lineamientos y Criterios Generales para la definición, identificación y medición de la pobreza” del Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL, 2015), publicados en el Diario Oficial de la Federación el 16 de junio de 2010. Asimismo, el CUIS forma parte integral de las Reglas de Operación del Programa Pensión para Adultos Mayores, para el ejercicio fiscal 2015, las cuales fueron publicadas en el Diario Oficial de la Federación el sábado 27 de diciembre del 2014, en su octava sección.

El instrumento aplicado en la investigación se centra en la información relacionada con los indicadores de ingreso corriente per cápita, rezago educativo promedio en el hogar, acceso a los servicios de salud, acceso a la seguridad social, calidad y espacios de la vivienda, acceso a los servicios básicos en la vivienda, y acceso a la alimentación. Consta de 39 preguntas de las cuales, 34 se enfocan a verificar el ingreso y seis tipos de carencia de conformidad con los criterios que da el CONEVAL para conocer la situación socioeconómica en la que se encuentran las personas. Las otras 5 preguntas también tienen que ver con condiciones socioeconómicas en cuanto a calidad de espacios en la vivienda y acceso a la alimentación..

3. RESULTADOS

Como puede apreciarse en la Tabla 1, entre los años 2013-2015 las carencias sociales de los hogares con adultos mayores beneficiados con el PPAM observaron una sustantiva disminución, exceptuando calidad y espacios en la vivienda. En lo que corresponde al ingreso corriente per cápita de los hogares encuestados, se observó un avance marginal durante el período de análisis, por lo que constituye un aspecto en el cual deberán redoblarse esfuerzos institucionales.

Tabla 1. Cambios en las dimensiones de la pobreza 2013-2015.

	2013	2015
Carencia por rezago educativo promedio en el hogar;	156	139
Carencia por acceso a los servicios de salud;	72	4
Carencia por acceso a la seguridad social;	173	0
Carencia por calidad y espacios de la vivienda;	35	39
Carencia por acceso a los servicios básicos en la vivienda;	137	42
Carencia por acceso a la alimentación;	164	9
Carencia por ingreso corriente per cápita	173	170
Ingreso por debajo de la Línea de bienestar	173	170
Ingreso por debajo de la Línea de bienestar mínimo	173	168

Nota: Los datos obtenidos en el año 2013 son del SIFODE y los del año 2015 son los recolectados por el cuestionario aplicado.

Fuente: Elaboración propia

En lo que corresponde a rezago educativo promedio en el hogar, se observó una disminución del 12.82 % entre estos dos años. De todas las carencias sociales analizadas ésta es la que tuvo un menor decremento, sin embargo constituye un avance sustancial por tratarse de adultos mayores.

La dimensión acceso a los servicios de salud observó una importante disminución, ya que del 100% de hogares que enfrentaban dicha carencia en el año 2013, al 2015 sólo el 5.5% continuaban aún sin acceso a dichos servicios. Asimismo se observó que el 97% de las 68 personas que salieron de esta carencia, fue debido a su incorporación al programa federal conocido como Seguro Popular de Salud, en tanto que el 3% restante fueron inscritos al Instituto Mexicano del Seguro Social. Este resultado permite visualizar que la política actual gobierno federal en materia sanitaria ha contribuido a mejorar la calidad de vida de la población adulta mayor.

En lo que corresponde al indicador de acceso a la seguridad social, la investigación arroja un avance sustancial, ya que se logró erradicar por completo dicha carencia, observándose que de los 173 adultos mayores que en 2013 se encontraban en dicho estatus, al 2015 todos habían salido ya de dicha situación. Lo anterior se logró gracias a la incorporación de los adultos mayores al PPAM. Otro hecho relevante de la investigación es el hallazgo de que el 30% de los hogares entrevistados perciben hasta dos pensiones por parte del PPAM. Ello debido a la presencia de dos adultos mayores en un mismo hogar, tendencia demográfica irreversible en nuestro país.

La dimensión calidad y espacios de la vivienda es la única carencia que observó un retroceso marginal, pues durante el período de análisis se incrementó en un 11.42 %. En contraste, la dimensión acceso a los servicios básicos en la vivienda observó una disminución sustancial del 69.34 %. De esta manera, durante el período de análisis, 95 de 137 hogares con adultos mayores lograron salir de dicha carencia.

Respecto a la variable acceso a la alimentación, el análisis arroja un importante avance en materia de seguridad alimentaria durante el período analizado, ya que de 164 hogares con adultos mayores que vivían con carencia alimentaria, al 2015 únicamente 9 hogares persistían en dicha condición. Esto representa una disminución del 94.51 % en esta carencia.

En cuanto a la carencia por ingreso corriente per cápita, comentamos ya que el avance es positivo aunque marginal, toda vez que la carencia de nivel de ingreso por debajo de la línea de bienestar se redujo poco menos del 2%, y el ingreso por debajo de la línea de bienestar mínimo en un 3%. Al respecto cabe mencionar que este resultado es congruente, toda vez que el PPAM lo que busca es elevar el bienestar general de sus beneficiarios a través de asegurarles un ingreso mínimo acompañado con otros apoyos de protección social.

4. CONCLUSIONES Y RECOMENDACIONES

En este trabajo se analizó la evolución durante el período 2013-2015 de las condiciones de bienestar de los hogares beneficiados del programa Pensión para Adultos Mayores (PPAM), mismos que en el año 2013 se encontraban en pobreza extrema en el municipio de Centro, Tabasco. Un hallazgo que se debe resaltar es la disminución sustancial en las carencias sociales, ya que en el año 2015 el 88.55 % de la población contaba con menos de tres carencias, hecho que contribuyó a la disminución de la intensidad de la pobreza. De esta manera, el PPAM logró erradicar la carencia de acceso a la seguridad social al otorgar al 100% de la población estudiada un esquema de seguridad social contribuyendo así a su bienestar mediante el aseguramiento de un ingreso mínimo, así como la entrega de apoyos de protección social a los adultos mayores.

Por otra parte, el Programa del Seguro Popular de Salud, mismo que brinda protección en salud a la población no derechohabiente mediante un seguro de salud, público y voluntario orientado a reducir los gastos médicos de bolsillo y fomentar la atención oportuna a la salud, logró sacar de la carencia por acceso a los servicios de salud al 97% de las personas investigadas, mismas que en el año 2013 se encontraban carentes de dicha protección. En lo que respecta a la carencia por rezago educativo, ésta logró disminuirse en un 12.82% lo que representa un logro importante del Instituto de Educación para Adultos de Tabasco, ya que si bien en cifras no constituye un avance tan contundente, toma relevancia toda vez que se trata de adultos mayores de 65 años, mismos que tuvieron la voluntad de estudiar en esta etapa de su vida. Por otro lado, la carencia por acceso a los servicios básicos de la vivienda muestra una disminución sustancial del 69.34%.

Por su parte, en la carencia por acceso a la alimentación se observó un avance importante ya que se logró una disminución del 94.51% en los hogares con esta privación. Aquí juegan un papel importante dos programas institucionales: el Programa de Apoyo Alimentario mismo que otorga hasta \$660.00 pesos bimestralmente para compra de productos de la canasta básica alimentaria en tiendas Diconsa, así como el Programa de Comedores Comunitarios, el cual busca mejorar las condiciones nutricionales de la población de niños y niñas de 0 a 11 años de edad, mujeres en gestación y lactantes, personas con alguna discapacidad y adultos mayores de 65 años.

Finalmente, los resultados de la investigación permiten afirmar que durante el período 2013-2015 los hogares con adultos mayores de 65 años en pobreza extrema han mejorado en general sus condiciones

socioeconómicas, al haber observado una disminución en sus carencias básicas, exceptuando el indicador de calidad y espacios en la vivienda. De esta manera, el 91.32 % de los 173 hogares de adultos mayores que se encontraban en el año 2013 en pobreza extrema lograron salir de ese estatus gracias a una importante disminución en sus carencias sociales.

Sin embargo, un dato contrastante es que en el 2015 el 97.1% de los hogares encuestados aún se encontraban en situación de pobreza alimentaria, esto debido a que sus ingresos continuaban estando por debajo de la línea de bienestar mínimo, lo que implica que seguían siendo insuficientes para adquirir la canasta básica alimentaria. Ello permite concluir que si bien con la aplicación del PPAM se ha logrado un avance sustancial en la disminución de las carencias sociales de los hogares con adultos mayores, es necesario acompañar este esfuerzo con otros programas institucionales, de manera tal que les permitan mejorar sus ingresos.

La investigación demuestra también que los hogares con adultos mayores están mejorando su nivel de bienestar no como consecuencia de cambios que ocurren de manera natural o automática a lo largo del tiempo, si no como resultado de políticas públicas que han logrado mejorar las condiciones socioeconómicas de la población vulnerable. En su conjunto, los resultados sugieren también que los hogares beneficiados del programa Pensión para Adultos Mayores logran reducir su pobreza a lo largo del tiempo, pero no lo suficiente para que logren escapar de la pobreza, ni para mejorar los ingresos de estas familias. Al respecto, se ha logrado un gran avance con la estrategia federal denominada Cruzada Nacional contra el Hambre apoyada con una adecuada focalización de los programas federales, ayudando esto a una reducción de la población en pobreza extrema en el plano nacional. Sin embargo, es recomendable mejorar la coordinación interinstitucional para erradicar de manera eficaz la pobreza en sus diferentes status.

Asimismo, es recomendable acompañar las políticas públicas con otros programas sociales que permitan a estas familias y a sus generaciones futuras generar ingresos por arriba de la línea de bienestar mínimo.

REFERENCIAS

- [1] Banco Mundial. (2001). Informe sobre el desarrollo mundial 2000/2001: lucha contra la pobreza, Washington: Banco Mundial. Recuperado de <http://siteresources.worldbank.org/INTPOVERTY/Resources/WDR/Spoverv.pdf>
- [2] CONAPO. (2015). Aspectos generales de los resultados de las proyecciones de población, Recuperado de http://www.conapo.gob.mx/es/CONAPO/Aspectos_Generales_de_los_resultados_de_las_Proyecciones_de_Poblacion.
- [3] CONEVAL (2015), Medición de la Pobreza, Glosario, Recuperado el 10 de Agosto de 2015, de <http://www.coneval.gob.mx/Medicion/Paginas/Glosario.aspx>
Enciclopedia de Economía (S.F). Recuperado de <http://www.economia48.com/spa/d/bienestar-economico/bienestar-economico.htm>.
- [4] CONEVAL (2015), Medición de la Pobreza, indicadores de carencia social, Recuperado de <http://www.coneval.gob.mx/Medicion/Paginas/Medici%C3%B3n/Indicadores-de-carencia-social.asp>.
- [5] Hernández, J. (1996). Evaluación de programas de salud. En Buena-Casal, Caballo & Sierra (1996). Manual de evaluación en Psicología clínica y de la salud. Madrid: Siglo XXI.
- [6] Hernández, R., Fernandez, C. y Baptista P. (2006). Metodología de la investigación. (4ª ed.). México: McGraw Hill Interamericana.
- [7] Garcia, J. F. (2015). Aproximación al estudio de la pobreza en México: Propuesta de política de estado contra la pobreza (1ª Ed.). México: UJAT.
- [8] Piketty, T. (2014). El capital en el siglo XXI. México: Fondo de Cultura Económica.
- [9] Székely, M. (2005). Pobreza y Desigualdad en México entre 1950 y el 2004. Serie Documentos de Investigación, 24. Secretaría de Desarrollo Social.

LA INTEGRACIÓN DE RESULTADOS DE LAS PUBLICACIONES CIENTÍFICAS A PARTIR DE LAS PRUEBAS DE SIGNIFICACIÓN EN EL CASO DE LA PRUEBA PARA LAS MEDIAS DE DOS POBLACIONES

Pedro A. Monterrey. Departamento de Matemáticas
Universidad del Rosario, Bogotá, Colombia.

ABSTRACT

The word Meta-analysis group a set of statistical procedures devoted to combine results from different papers or scientific reports. If studies considered in meta-analysis obtained their conclusions applying significance test, the random nature of P-values gives an additional argument in favor of the relevance of meta-analysis. One way to combine results from different papers is calculating a combined P-value, despite its rationale this procedure is not widely used and graphical representations combined with analytical procedures based on fixed or random effects models are mainly considered in meta-analysis. In this paper P-value probability distribution for t-test was used to analyze homogeneity of results considered in meta-analysis and to quantify the evidence contained in studies in favor of the null-hypothesis

KEYWORDS: Meta-analysis. P-value distribution. T-test two populations.

RESUMEN

Por meta-análisis se entiende un conjunto de procedimientos estadísticos dedicados a combinar los resultados de diferentes artículos o reportes científicos. Si los resultados considerados en el meta-análisis obtuvieron sus resultados utilizando pruebas de significación, la naturaleza aleatoria del valor P brinda un argumento adicional a favor de la relevancia del meta-análisis. Una forma de combinar los resultados de diferentes artículos es calcular un valor P combinado, a pesar de su racionalidad este procedimiento no es utilizado frecuentemente, realizándose usualmente el análisis de los datos utilizando una combinación de procedimientos gráficos y analíticos basados en modelos de efectos fijos o aleatorios. En este artículo la distribución del valor P para la prueba t fue utilizada para analizar la homogeneidad de los resultados considerados en el meta-análisis y para cuantificar la evidencia contenida en los estudios a favor de la hipótesis nula.

1. INTRODUCCIÓN

Las revisiones sistemáticas de la literatura constituyen una etapa de búsqueda y selección de las publicaciones y reportes de investigación que abordan un tema de interés en la investigación. Estas revisiones constituyen, como señalan Avilés y cols [1], una primera aproximación a la solución del problema del gran volumen y dispersión de la información disponible. Con las revisiones sistemáticas se pretende sistematizar los conocimientos acumulados sobre un tema, sistematizando la información existente e identificando las características principales de los artículos y reportes publicados hasta el momento de su realización. Inferir conclusiones de estas agrupaciones de artículos pudiera resultar en la obtención de conclusiones errónea, son múltiples los factores que se reconoce influyen sobre la calidad de la información recolectada. Desde el punto de vista estadístico se han desarrollado procedimientos específicos de análisis para realizar esta integración, así una vez sistematizada la información se realiza un meta-análisis. El término meta-análisis se utiliza para agrupar un grupo de métodos cuantitativos de integración de la literatura científica. La acepción más reconocida del término se refiere solamente a los métodos utilizados directamente para contrastar y combinar, mediante técnicas estadísticas, los resultados de diferentes estudios.

En muchos artículos científicos, las conclusiones se basan en diseños de muestreo que determinan muestras aleatorias las que se analizan utilizando pruebas de hipótesis. El carácter aleatorio del valor P que se utiliza en esos análisis determina, según las consideraciones metodológicas de Neyman y Pearson y que fueron discutidas en Monterrey [4] que los procedimientos de pruebas de hipótesis arrojan resultados válidos solamente en el contexto de múltiples repeticiones del experimento. Esta situación potencia la validez del meta-análisis como procedimiento de valoración de las evidencias científicas. Una primera forma de combinar resultados de diferentes estudios es combinar los valores P obtenidos de la aplicación de pruebas de hipótesis semejantes para abordar el mismo problema, una primera opción para esto fue introducida por

Fisher en 1925, pero posteriormente se han desarrollado diferentes procedimientos con mayores o menores ventajas, detalles de estas opciones pueden identificarse en Sungho y cols. [7] o en Koziol y Perlman [3]. La combinación de valores P, aunque es una posibilidad, no es muy explotada al realizar los meta-análisis, lo usual es combinar procedimientos gráficos con procedimientos analíticos específicos basados en modelos de efectos fijos, efectos aleatorios o siguiendo procedimientos bayesianos, detalles de estas formas de proceder pueden encontrarse en Fleiss [2].

La naturaleza aleatoria del valor P utilizado para el análisis de las pruebas de significación y el que, cuando la hipótesis nula sea cierta, su distribución es estocásticamente mayor o igual que la distribución uniforme sobre el intervalo [0,1] tal y como se analiza en Monterrey [5] permitiría complementar los análisis de homogeneidad de los estudios que se integran en el meta-análisis, así como analizar la validez de la hipótesis nula. El objetivo del presente trabajo es utilizar la distribución del valor P, en el caso de las pruebas para las medias de dos poblaciones, para analizar la homogeneidad de los estudios y sus evidencias a favor o en contra de la hipótesis nula o de ausencia de efecto.

2. EL ANÁLISIS CONJUNTO DE EVIDENCIAS PROVENIENTES DE DIFERENTES ESTUDIOS

Para una familia de k estudios realizados para comparar las respuestas a dos tratamientos, aplicados independientemente con una respuesta continua: $X_T \sim n(\mu_T, \sigma^2)$, representa la respuesta al tratamiento experimental y $X_R \sim n(\mu_R, \sigma^2)$ la respuesta al tratamiento de control en el estudio $i, i=1, \dots, k$. Denotando $n(\mu, \sigma^2)$ la distribución normal con los parámetros correspondientes. En un meta-análisis el interés es combinar las evidencias contenidas en cada uno de los k estudios para determinar si las medidas de efecto, en este caso representadas por las medias, son iguales o diferentes. Partiendo de las características, en general heterogéneas de los estudios que se combinan, el parámetro de interés, la medida de efecto a considerar, es $\theta = (\mu_T - \mu_R)/\sigma$. El objetivo del meta-análisis es utilizar de manera combinada la información de todos los estudios para determinar la validez de las hipótesis $H_0: \theta = 0$ $H_A: \theta \neq 0$.

En el estudio i $\bar{X}_{T,i}$, $\bar{X}_{R,i}$ representarán, respectivamente, las medias muestrales de la medición que representa la respuesta al tratamiento, n y m denotarán los respectivos tamaños de las muestras y $S_{T,i}^2$ y $S_{R,i}^2$ las varianzas muestrales correspondientes. En el estudio i , $S_{P,i}^2 = (n+m-2)^{-1} [(n-1)S_{T,i}^2 + (m-1)S_{R,i}^2]$ representará la varianza combinada y $\hat{\sigma}^2$ el promedio de las varianzas ponderadas en los diferentes estudios, estadístico que es un estimador insesgado de σ^2 . Para cada i $\hat{\theta}_i = \sigma^{-1}(\bar{X}_{T,i} - \bar{X}_{R,i})$ es el estimador del efecto del tratamiento T en el estudio i ; aproximándose, según indica Fleiss [2] σ^{-2} por $S_{P,i}^2$. Es claro que $V(\hat{\theta}_i) = n_i^{-1} + m_i^{-1}$, denotando $V(\cdot)$ la varianza del estimador. En el modelo de efectos fijos se utiliza el estimador combinado del efecto θ , $\hat{\theta} = (\sum_i \omega_i)^{-1} (\sum_i \omega_i \hat{\theta}_i)$ con $\omega_i = n_i m_i (n_i + m_i)^{-1}$. Como $V(\hat{\theta}) = (\sum_i \omega_i)^{-1}$ con el estimador $\hat{\theta}$ se construye el estadígrafo $Z = \hat{\theta} / \sqrt{V(\hat{\theta})}$ que, aproximadamente se distribuye $n(0,1)$ bajo H_0 y que se utiliza para realizar la prueba de dos colas para analizar $H_0: \theta = 0$ $H_A: \theta \neq 0$, combinando así los resultados de los diferentes estudios. El procedimiento de combinación es válido si existe homogeneidad entre estudios.

Una propuesta para caracterizar el aporte individual de cada estudio al análisis conjunto y adicionalmente caracterizar si la integración de las evidencias de los diferentes reportes es concordante o discordante con la hipótesis nula se puede obtener utilizando la distribución de probabilidad del valor P para analizar las hipótesis. El primer paso de la propuesta es evaluar las hipótesis $H_0: \theta = 0$ $H_A: \theta \neq 0$ en cada uno de los estudios utilizando el estimador $\hat{\theta}_i$, así para $i=1, \dots, k$ se considera el estadígrafo $\hat{\theta}_i = \sigma^{-1}(\bar{X}_{T,i} - \bar{X}_{R,i})$ como medida de efecto del tratamiento experimental en el estudio i y el estadígrafo de prueba:

$$Z_i = \hat{\theta}_i \sqrt{\frac{n_i m_i}{n_i + m_i}} \quad (1)$$

Para θ número real, $i=1, \dots, k$; $Z_i \sim n(\Delta_i, 1)$ con $\Delta_i = \theta \left(\frac{1}{n_i} + \frac{1}{m_i} \right)^{\frac{1}{2}}$ y según lo indicado en Monterrey [5] el valor P de cada prueba individual es una realización de la variable aleatoria $\Pi_i = 2 \Phi(-|Z_i|)$. Su distribución de probabilidad se determina por:

$$(0 < x < 1) \quad F_{\Pi_i/\theta}(x) = P(\{|Z_i| > -z_{x/2}\}) = 1 - \Phi(z_{1-x/2} - \Delta_i) + \Phi(z_{x/2} - \Delta_i) \quad (2)$$

Si H_0 es cierta Π_i se distribuye uniforme en el intervalo $[0,1]$.

Realizando las pruebas determinadas por el estadígrafo Z_i , definido en (1), para cada estudio, utilizando la misma estimación de la varianza ($\hat{\sigma}^2$), que es una estimación de la varianza combinando las evidencias de todos los estudios, en lugar de las varianzas combinadas de cada uno de ellos individualmente, que es lo que indica el procedimiento usual se podría controlar el parámetro de molestia y disponer de distribuciones del valor P comparables; a partir de lo cual se podría definir una distancia entre las probabilidades de los valores P observados en cada estudio y los esperados si H_0 fuera cierta. Esto puede hacerse de la siguiente forma:

$$\delta = \sum_{i=1}^k (F_{\Pi_i/\hat{\theta}}(p_i) - p_i)^2 \quad (3)$$

Donde las funciones de distribución del valor P se calculan asumiendo que la estimación $\hat{\theta}$ es el valor del parámetro y representando en (3) p_i el valor observado del valor P en cada estudio.

Considerando (3) como una medida de lo discrepantes que son los valores P observados de los que se observarían si H_0 fuera cierta es posible cuantificar el aporte de cada estudio a ese valor, determinándose de esta forma los estudios que no tienen un comportamiento semejante a los otros y por tanto aquellos que no son consistentes con el supuesto de homogeneidad:

$$\delta_i = \frac{F_{\Pi_i/\hat{\theta}}(p_i) - p_i}{\delta(p_1, \dots, p_k)} \quad (4)$$

3. UN EJEMPLO DE APLICACIÓN

Tomando, por ejemplo algunos de los estudios del meta-análisis realizado por Munford y cols [6] y que fue referido por Fleiss [2] en el cual se pretendía determinar el efecto de la psicoterapia sobre el número de días de hospitalización de los pacientes. La Tabla 1 muestra los resultados de 9 de esos estudios

Tabla 1. Resultados de los estudios del meta-análisis

no	Estudio	Psicoterapia			Control		
		n	media	DE*	n	media	DE*
1	Davis	13	5	4.7	13	8.5	3.8
2	Archuletta	248	7.49	5.7	267	6.9	3.9
3	Florell	30	4.9	1.7	50	6.1	2.3
4	Gruen	35	22.5	3.4	35	24.9	10.7
5	Hart	20	12.5	1.5	20	12.3	1.7
6	Hill	10	3.4	0.9	10	3.19	0.8
7	Jacobson	13	4.9	1.1	14	5.5	0.9
8	Johnson	13	3.6	0.8	14	3.9	0.9
9	Wilson	8	6.5	0.8	8	7.38	1.4

* DE: Desviación Estándar

A partir de los valores presentados en la Tabla 1 se calcula $\hat{\theta} = -0,062$ y $Z = 0,0000$ por lo que el valor $p = 1,000$ y no se rechaza $H_0: \theta = 0$. La prueba Q para el análisis de la homogeneidad presenta un valor $Q_{obs} = 0,095$, como el estadígrafo de la prueba tiene distribución ji-cuadrado con 8 grados de libertad, $p = 1,000$ por lo que no se rechaza la hipótesis de homogeneidad del efecto entre estudios. Este sería el análisis usual en el meta-análisis para integrar los diferentes artículos y responder a la pregunta de si la psicoterapia influye sobre el tiempo de hospitalización.

Aplicando la propuesta de análisis utilizándola distribución de probabilidad del valor p. De la Tabla 1 se calcula también que $\hat{\delta}^2 = 12.83$. La Tabla 2 presenta los valores definidos en (3) y (4). De la Tabla 2 se obtiene que $\delta = 0,0046$, lo que indica que los valores P observados son concordantes con los de la distribución uniforme lo cual es una evidencia a favor de $H_0: \theta = 0$ y complementa el análisis tradicional. Adicionalmente se aprecia que, excepto en estudio de Archuletta, el resto de los artículos no tienen grandes discrepancias en su aporte a las diferencias entre las probabilidades observadas y las esperadas suponiendo H_0 . La discrepancia del valor P del artículo 2, que no es muy elevada en términos absolutos, se debe fundamentalmente a lo elevado de su tamaño de muestra comprado con los utilizados por los otros autores.

Tabla 2. Análisis de la Distribución del valor P

no	Estudio	probabilidad acumulada según (2) para $\theta = 0,062$	Probabilidad acumulada según uniforme	aporte individual a la distancia
1	Davis	0.0139	0.0128	0.000
2	Archuletta	0.1282	0.0619	0.945
3	Florell	0.1617	0.1470	0.046
4	Gruen	0.0066	0.0051	0.001
5	Hart	0.8626	0.8599	0.002
6	Hill	0.8967	0.8957	0.000
7	Jacobson	0.6679	0.6637	0.004
8	Johnson	0.8301	0.8279	0.001
9	Wilson	0.6260	0.6233	0.002

REFERENCIAS

- [1] Aviles R., Morales M., Sao A. y Canedo R. (2004): La Colaboración Cochrane en Cuba: Parte VII. Los meta-análisis.: aproximaciones útiles para su comprensión. ACIMED [online], 12, [citado 2016-06-15], pp. 1-1. Disponible en: <http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S1024-94352004000400004&lng=es&nrm=iso>. ISSN 1024 - 9435.
- [2] Fleiss J.L. (1993): The statistical basis of meta-analysis. *Statistical Methods in Medical Research*, 2, 121-145
- [3] Kosiol J. and Perlman M. (1978): Combining Independent chi-squared test. *Journal of the American Statistical Association*, 73:753-763
- [4] Monterrey P. (2012): $p < 0,05$, ¿Criterio mágico para resolver cualquier problema o leyenda urbana? *Universitas Scientiarum*. *Universitas Scientiarum*, 17, 203-215
- [5] Monterrey P. (2016): Caracterización de los valores P de las pruebas paramétricas en poblaciones normales. *Investigación Operacional*, 37, 38-47
- [6] Mumford E., Schlesinger H., Glass G., Patrick C. and Cuedon T: (1984) A new look at evidence about reduced cost of medical utilization following mental health treatment. *The American Journal of Psychiatry*, 141, 1145-1158
- [7] Sungho W., Morris N., Qing L. and Elston R. (2009): Choosing an optimal method to combine P-values. *Statistics in Medicine*, 28, 1537-1553

MÉTODOS ALTERNATIVOS A LAS TÉCNICAS DE RESPUESTA ALEATORIZADA PARA ESTUDIAR COMPORTAMIENTOS SENSIBLES EN ENCUESTAS DE SALUD

Alberto Baena*, Beatriz Cobo* & María del Mar Arcos**

*Universidad de Granada, España

**Hospital de La Paz

ABSTRACT

The present paper includes a review of several techniques of indirect interview developed as an alternative to the randomized response technique and the use of them in real surveys for studying sensitive characteristics related to human health.

KEYWORDS: Randomized response, Sampling, Desirability social bias, Item count techniques

RESUMEN

En este trabajo se hace una revisión de las diversas técnicas de entrevista indirecta surgidas como alternativa a la técnica de respuesta aleatorizada y de su utilización en encuestas reales destinadas a estudios de características sensibles relativas a la salud humana.

1. INTRODUCCIÓN

La recogida de datos mediante encuestas personales relativas a variables sensibles tales como abortos, uso de drogas, enfermedades venéreas, violencia de género, acoso sexual, relaciones extramatrimoniales, etc, es un problema muy importante en la investigación de encuestas. Las técnicas de respuesta aleatoria fueron introducidas por Warner en 1965 con el objetivo de proteger la privacidad de los encuestados en el estudio de temas sensibles y de este modo evitar el sesgo de respuesta. Realizar estimaciones válidas de la prevalencia o frecuencia de estos comportamientos sensibles es difícil por el hecho de que las personas dan valores mucho más bajos de estos comportamientos socialmente inapropiados. Hay muchos estudios que evidencian grandes sesgos debido a la sensibilidad de estos temas, dependiendo del formato de las preguntas, el modo de recogida de datos, las características de entrevistados y el encuestado, etc.

Los investigadores han intentado combatir estos sesgos de respuesta que sobreestiman o subestiman los valores de las variables sensibles dependiendo de la deseabilidad del comportamiento en cuestión, de una gran variedad de formas. Muchas de ellas están enfocadas a proporcionar métodos que garanticen al encuestado y alto grado de confidencialidad, de forma que al aumentar su anonimato los encuestados aumenten tanto su disposición a responder como la veracidad de sus respuestas.

Entre estos métodos la más conocida es la técnica de respuesta aleatorizada introducida por Warner en 1965. Esta técnica consiste en la utilización de un mecanismo aleatorio por medio del cual se selecciona una de dos preguntas complementarias: ¿pertenece al grupo con la característica A ? o ¿pertenece al grupo que no tiene la característica A ?, donde A es la característica sensible de interés. El entrevistado contestará sí o no y el entrevistador no tiene la posibilidad de saber qué pregunta contestó el entrevistado. Mediante esta técnica se obtiene una cooperación mayor por parte de los encuestados y reduce su motivación para falsear las respuestas dando lugar a estimadores que son más precisos en comparación a respuesta directa.

Este trabajo de Warner originó una enorme literatura sobre las técnicas de respuesta aleatoria, entre los que destacan autores como Horvitz et al. (1967), Greenberg et al. (1969), Kuk (1990), Mangat y Singh (1990), Mangat (1994), Kim y Warde (2004), Chang et al. (2004), Gjestvang y Singh (2006), Chaudhuri (2010), Mehta et al. (2012), Gupta et al. (2013), entre otros. La respuesta aleatorizada ha sido utilizada en muchos ámbitos, sobre todo en las referidas a la prevalencia de prácticas inmorales e ilegales. Pero estas técnicas tienen dificultades y limitaciones como son:

- La exigencia de comprensión de los encuestados hacia el cuestionario y el tiempo empleado en la realización del mismo,

- la necesidad de realizar entrevistas cara a cara con el entrevistado para generar respuestas válidas
- el pensamiento por parte del encuestado de que el procedimiento puede estar trucado.

Medio siglo después dichas técnicas continúan utilizándose, aunque otros métodos alternativos han sido propuestos, entre los que destacan las técnicas indirectas como alternativas a la respuesta aleatorizada. Estas técnicas son:

1. La Técnica de Conteo de Items.
2. La Técnica Nominativa.
3. El Método de las Tres Cartas.
4. Los Modelos No Aleatorizados.
5. El Estudio con Preguntas Negativas.

En este trabajo se van a presentar estas distintas técnicas indirectas como alternativas a la respuesta aleatorizada.

2. TÉCNICA DE CONTEO DE ITEMS

Esta técnica es la más conocida entre las técnicas indirectas, cuya función se basa en estimar la proporción de personas que tienen una característica sensible, introducida por Raghavarao y Federer (1979) como alternativa a la respuesta aleatorizada, llamándola *Block Total Response*. Dicha técnica surgió como alternativa a la respuesta aleatorizada y se basa en una pregunta directa y a su vez anónima buscando una respuesta a una cuestión sensible o a una serie de v cuestiones sensibles, que se pretenden estimar. Dichas respuestas deben de ser obtenidas para todos los miembros de la población o para una submuestra, tanto para una como para más de dos cuestiones. La intención es contar el total de k respuestas afirmativas de las cuestiones, tanto sensitivas como no sensitivas, que son aplicadas a cada encuestado. Además, cada uno de ellos posee diferentes cuestiones, ya que hay b grupos diferentes de preguntas diseñadas con el propósito de estimar el diseño de bloque experimental que se va a realizar, tanto un diseño de bloque completo (SB) o incompleto (BIB).

Para ello, el bloque proporcionado es asignado aleatoriamente a cada encuestado y los bloques tienen el mismo número o aproximado de encuestados, obteniéndose así, tanto estimadores de la proporción poblacional como de la media para cada una de las cuestiones, siendo imposible determinar una respuesta individual a una pregunta en particular.

Más tarde, Miller (1984) desarrolló y demostró empíricamente esta técnica también denominada como *tecnica de conteo de items* o *técnica de lista experimental*. Para ello se basa en que cada individuo posee un cuestionario con una lista de elementos, donde el investigador pide al individuo enumerar los elementos de dicho cuestionario aplicables a él/ella. Inicialmente, este autor, proporciona una versión más simple donde propone que dos submuestras son elegidas de manera arbitraria, denominadas submuestra A y submuestra B . En la submuestra A es presentada una lista con cuatro elementos, donde uno de ellos puede tener un comportamiento sensible hacia su persona. En la submuestra B se presenta una lista con los mismos elementos inocuos, aunque el elemento sensible del cuestionario es eliminado. A partir de ahí, se pueden comparar las respuestas de ambas muestras con la diferencia de ambas medias, \bar{X}_{4A} y \bar{X}_{3B} , respectivamente, que proporciona la estimación. Por tanto, sea P quien denota el parámetro que se quiere estimar, un estimador insesgado para este parámetro es:

$$\hat{p} = \bar{X}_{4A} - \bar{X}_{3B}$$

Aunque la varianza asociada es bastante alta, por lo que se introdujo una versión de dobles listas para solucionar este problema, aumentando así la eficacia del estimador ya que se generan dos estimaciones en una sola estimación final con el promedio. Así pues, aplicando la operación anterior se obtiene el primer estimador, mientras que el segundo estimador es proporcionado con dos nuevas listas con las mismas características que el estimador anterior (X_4 e X_3), es decir, dos cuestionarios con tres elementos inocuos y uno de ellos tiene añadido un elemento sensible, (Y_4 e Y_3). Por tanto, los encuestados de la submuestra A responden a los cuestionarios X_4 e Y_3 siendo sus medias respectivas \bar{X}_{4A} e \bar{Y}_{3A} , y los encuestados de la submuestra B responden a los cuestionarios Y_4 e X_3 siendo sus medias \bar{Y}_{4B} e \bar{X}_{3B} , respectivamente. Por tanto, se calcula el promedio con sus medias respectivas de la siguiente forma:

$$E(P) = \hat{p} = 1/2[(\bar{X}_{4A} - \bar{X}_{3B}) + (\bar{Y}_{4B} - \bar{Y}_{3A})]$$

Con este método, a comparación de la versión más simple, su varianza se ve reducida a la mitad. Ha sido utilizada por otros autores como Kuklinski et al. (1996) y Corstange (2009).

Aunque una revisión de Miller (1984), tuvo como consecuencia el método definitivo, agrupando datos sobre una característica binaria sensible. Para ello, se realiza una elección de individuos aleatoria en

dos muestras, una de control y otra de tratamiento. Al grupo de control, se le proporciona una lista con “G” elementos inocuos. Por otra parte, al grupo de tratamiento, se le proporciona una lista con “G+I” elementos, con las mismas preguntas inocuas añadiendo una sensible que es la que se quiere estimar. En ambos grupos, el individuo debe proporcionar el número de elementos que son aplicables a él o ella.

Sea $U_j \in \{0,1\}$ ($j = 0, \dots, G$) e $Y \in \{0,1\}$ para un elemento inocuo y sensible respectivamente, se define $T = 0$ si el individuo es asignado al grupo de control, mientras que si es asignado al grupo de tratamiento el valor asignado es $T = 1$. Además, sea $Z = U_1 + \dots + U_G$ la respuesta requerida para el grupo de control S , mientras que el grupo de tratamiento es instruido a dar una respuesta $Z + Y$. Se supone que en el grupo de control, al poseer elementos inocuos, el anonimato de los encuestados está bien protegido. Sin embargo, esto puede fallar cuando se aplica el elemento sensible en el grupo de tratamiento. En este caso, una respuesta del grupo de tratamiento implica que $Y = 1$.

Si se asume una muestra aleatoria simple de n individuos extraída e i como el i -ésimo individuo correspondiente a U_G, Y, T, Z, S por $U_{ij}, Y_i, T_i, Z_i, S_i$ respectivamente. Para denotar la proporción de personas que tienen $Y = 1$ para $\pi_0 = 1 - \pi_1$ y se define $\pi = (\pi_0, \pi_1)^t$, se puede estimar π_1 :

$$\hat{\pi}_1 = \frac{1}{n_T} \sum_{i=1}^n T_i S_i - \frac{1}{n_C} \sum_{i=1}^n (1 - T_i) S_i$$

siendo $n_T = \sum_{i=1}^n T_i$ y $n_C = n - n_T$. Aunque desafortunadamente, este método puede producir que el estimador tenga valores negativos o mayores que uno, además puede que todos los “G” elementos puedan ser aplicables a un individuo, o por el contrario que no sea aplicable ninguno, provocando que el anonimato de los individuos no quede bien protegida. Con esta técnica de conteo de ítems, se han llevado a cabo algunos test muy conocidos como el consumo de drogas llevado a cabo por Miller et al. (1986) donde fueron obtenidas estimaciones para el consumo de heroína para gente joven, que reflejaron estimaciones más altas por la técnica de conteo de ítems que por métodos directos.

En esta dirección, Droitcour et al. (1991) llevó a cabo un proyecto con el objetivo de estimar la prevalencia acerca de las infecciones del virus “VIH” en EEUU, sugiriendo que es muy importante que los encuestados entiendan bien el procedimiento desempañado y que se requiere una buena selección de los elementos no sensibles de los cuestionarios. Además, si el elemento sensible no es muy conocido, la técnica de conteo de ítems no es un buen método de selección.

En definitiva, si se procede a realizar una comparación de la *técnica de respuesta aleatorizada* y de la *técnica de conteo de ítems*, esta última tiene como ventaja que el procedimiento es realmente simple, además de poseer una carga moderada cognitiva al encuestado, aumentando la capacidad de cumplir con el cuestionario por parte del entrevistado y proporcionar autoinformes más honestos. La desventaja está en que la técnica de conteo de ítems tiene un poder estadístico bajo, es decir, errores estándar más grandes con el mismo tamaño de la muestra que la respuesta aleatorizada, contrastado por Coutts y Jann (2011).

Para solucionar los problemas de privacidad de la técnica de conteo de ítems que se han comentado anteriormente, Chaudhuri y Christofides (2007) proponen una serie de mejoras y revisiones con dos y tres muestras independientes respectivamente, ya que en la técnica de conteo de ítems todos los elementos o ninguno pueden ser aplicados a un individuo. Primeramente, da una revisión con dos muestras que está pensada para que el individuo que realiza el segundo cuestionario, ósea de la segunda muestra (donde se encuentra el elemento sensible), conteste al menos un elemento afirmativamente que se le presenta. Si esto no ocurriera, dieron una revisión para tres muestras, para mejorar este aspecto de privacidad.

También, Chaudhuri y Christofides (2007) propusieron una técnica de conteo de ítems para estimar características sensibles cuantitativas, por ejemplo el número de abortos inducidos, el número de veces que se realiza una actividad ilegal, la cantidad de dinero no declarado, etc. Esta técnica fue extendida y desarrollada por Trappmann et al. (2014) denominada *técnica de suma de ítems*, análoga a la *técnica de conteo de ítems*, utilizando dos submuestras independientes escogidas al azar, aunque los elementos que se le presentan en los cuestionarios son cuantitativos.

Por último, Hussain et al. (2012) propone una alternativa a la *técnica de suma de ítems* sin la necesidad de utilizar dos submuestras, es decir, solo se toma una muestra de individuos que realizan el cuestionario.

3. TÉCNICA NOMINATIVA

Esta técnica fue introducida por Miller (1985) pretendiendo que los individuos no revelen su propia información sensible, sino que revelen el número de personas cercanas o conocidas de él/ella que tienen un elemento sensible propuesto. Esto permite al investigador obtener una estimación de individuos que tienen un atributo sensible en una comunidad concreta. Dicha técnica, puede ser considerada como un

muestreo de red, es decir, utiliza la respuesta de un individuo sobre otros para aumentar los participantes. Esta forma de realizar el estudio fue introducida por Thompson (1992), y más tarde desarrollada por Thompson y Seber (1996) y Chaudhuri (2000). Para esto un requisito básico es la buena voluntad de proporcionar información por parte del encuestado sobre otros miembros de su comunidad.

Para dicho método, se parte de una población de tamaño N , siendo r_{ij} para i tomando el valor $r_{ij} = 1$ si el j -ésimo encuestado responde que el i -ésimo miembro de la muestra tiene la característica sensible y $r_{ij} = 0$ en otro caso.

Claramente, $\sum_{i=1}^N r_{ij}$ representa el número de individuos respondidos por el j -ésimo encuestado quienes tienen la característica sensible, al igual que $\sum_{j=1}^N r_{ij}$ representa el número de veces que el i -ésimo miembro de la población es dado como perteneciente a la característica sensible. Finalmente, $\sum_{i=1}^N \sum_{j=1, i \neq j}^N r_{ij}$ proporciona el número de personas que poseen la característica sensible. Por tanto, el total de personas en una comunidad con la característica sensible viene dada por:

$$T = \sum_{i=1}^N \sum_{j=1, i \neq j}^N (r_{ij} / \sum_{k=1}^N r_{ik})$$

donde $r_{ij} / \sum_{k=1}^N r_{ik}$ es la cantidad de respuestas que son tomadas como 0 si $\sum_{k=1}^N r_{ik}$ es igual a cero. Además, sea A_j el número de individuos que tienen la característica sensible a través del j -ésimo encuestado y sea B_j el número de individuos cercanos proporcionados por el j -ésimo encuestado que poseen la característica sensible, y por último, sea $x_j = A_j / (1 + B_j)$, el estimador insesgado para la proporción poblacional y para la población total vienen dados, respectivamente, por:

$$\hat{\theta} = \frac{1}{n} \sum_{j=1}^n x_j, \quad t = N\hat{\theta}.$$

Esta técnica ha tenido una serie de extensiones entre las que destacan la proporcionada por Chaudhuri y Christofides (2008), donde se asume una población de población finita. En esta técnica, se valoran las respuestas positivas a una característica sensible que ofrece una serie de personas sobre otras en una población definida. Se puede dar el caso, de que el encuestado no proporcione candidatos o que el total de ellos no proporcione información. Para este caso, Thompson (1992) y más tarde desarrollada por Thompson y Seber (1996), Chaudhuri (2000) y Chaudhuri et al. (2004). En esta técnica, un encuestado debe proporcionar si algún vecino suyo de su comunidad tiene una característica sensible que se ha presentado y si es así, todos los miembros de la comunidad son seleccionados para realizar el cuestionario. Aunque un problema de dicha técnica, es que puede ser que la muestra sea muy grande, y por lo tanto, costoso.

En definitiva, la *Técnica Nominativa* no ha sido utilizada en muchos casos, Kutnik et al. (2007) propusieron que esta técnica podría resultar interesante para estimar el número de personas que son víctimas del tráfico humano, aunque estos autores no han llegado a utilizarla. Por el contrario, si ha sido utilizada por John et al. (2010) para evaluar problemas acerca de la conservación de los recursos naturales.

4. EL MÉTODO DE LAS TRES CARTAS

Este método, introducido por Droitcour et al. (2001), es muy simple y requiere de tres muestras, las cuales deben de ser independientes y se puede llevar a cabo con dispositivos tales como cajas y cartas. Se va a introducir dicha técnica con un ejemplo, para que se vea fácilmente su procedimiento. Por tanto, partiendo de tres cajas, *Caja 1*, *Caja 2* y *Caja 3*, en cuyo interior de cada una de las tres cajas, poseen diferentes categorías de respuesta, de tal forma que en una de ellas hay incluido un elemento sensible, con otras categorías de respuesta que son inocuas. En las cajas restantes, solo hay una categoría de respuesta inocua. Si se asigna a A la categoría de respuesta que es sensible, y B, C y D las categorías inocuas, entonces el procedimiento que se lleva a cabo es el siguiente:

A la primera muestra se le presentan los siguientes caracteres:

- Caja 1: Pertenezco a B .
- Caja 2: Pertenezco a A, C o D .
- Caja 3: Pertenezco a algún otro grupo que no está en la *Caja 1* o *Caja 2*.

A la segunda muestra se le presentan los siguientes caracteres:

- Caja 1: Pertenezco a C .
- Caja 2: Pertenezco a A, B o D .
- Caja 3: Pertenezco a algún otro grupo que no está en la *Caja 1* o *Caja 2*.

A la tercera muestra se le presentan los siguientes caracteres:

- Caja 1: Pertenezco a D .

- Caja 2: Perteneczo a A, B o C .
- Caja 3: Perteneczo a algún otro grupo que no está en la *Caja 1* o *Caja 2*.

Una vez que se ha presentado las diferentes categorías de respuesta impuestas para cada muestra, el participante debe proporcionar el número de caja que es aplicable a él/ella. Si se asume que los encuestados pertenecientes a las categorías inocuas B, C y D , tienen proporciones poblacionales π_B, π_C y π_D respectivamente, entonces sus estimadores poblacionales serán $\hat{\pi}_B, \hat{\pi}_C$ y $\hat{\pi}_D$. Por ejemplo, en la primera muestra se puede estimar la proporción de individuos que tienen una característica sensible (A) propuesta, utilizando el número de personas que ha seleccionado la *Caja 2*. Así pues, sea $\hat{\pi}_{ICD}$ la proporción de individuos que ha seleccionado la *Caja 2*, se obtiene el siguiente estimador para la proporción poblacional de individuos que poseen la característica sensible que se ha propuesto:

$$\hat{\pi}_{IA} = \hat{\pi}_{ICD} - \hat{\pi}_C - \hat{\pi}_D.$$

Del mismo modo, se pueden obtener otros dos estimadores de la proporción poblacional con las otras dos muestras. Por tanto, se puede combinar los tres estimadores para obtener una media del estimador, aunque se debe tener claro, que tiene que haber una asignación óptima de muestras, y que A, B, C y D deben ser excluyentes mutuamente. Claro está, estas condiciones crean dificultades al investigador para encontrar tres rasgos exclusivos que no sean sensibles para B, C y D , y a su vez, mutuamente excluyentes con la categoría sensible A .

Los autores que propusieron dicha técnica, Droitcour et al. (2001) llevaron a cabo una investigación donde para estimar la proporción de individuos que residían ilegalmente en EEUU (*Grupo A*). Mientras que las categorías inocuas, eran (1) personas que tenían una tarjeta verde válida para residir publicada por el gobierno estadounidense (*Grupo B*), (2) personas estadounidenses (*Grupo C*) y (3) personas que están estudiando o con un visado de trabajo en EEUU. Obviamente, para estimar la proporción de personas que consume drogas no serviría dicha técnica.

5. MODELOS NO ALEATORIZADOS

Otra técnica para solventar los problemas que poseen las técnica de respuesta aleatorizada son los *modelos no aleatorizados*. Esta técnica fue recomendada por Tian et al. (2007), Yu, Tian y Tang (2008) y Tan, Tian y Tang (2009) para facilitar el estudio de un rasgo sensible, denominado A , para proteger el anonimato de una serie de individuos de una comunidad dada.

Así pues, sea i un individuo de la población N , entonces $y_i = 1$ si la persona i -ésima tiene la característica A , mientras que $y_i = 0$ si la persona i -ésima tiene la característica complementaria a A (A^c). Por otro lado, la variable $x_i = 1$ si el encuestado i -ésimo tiene la característica inocua B y $x_i = 0$ si tiene la característica complementaria a la característica B . A continuación, se presenta un ejemplo para esta variable x :

Si el individuo nace entre los meses de agosto y diciembre toma el valor 1, ya que es una respuesta afirmativa a la pregunta inocua dada. Si por el contrario, el individuo nace entre los meses de enero y julio x toma el valor 0, es decir, su respuesta es negativa a que haya nacido entre agosto y diciembre.

Además, Tian et al. (2007), Yu, Tian y Tang (2008) y Tan, Tian y Tang (2009), proporcionan una extensión para un diseño muestral general, no solo para un muestreo aleatorio simple con reemplazamiento.

Más tarde, Christofides (2009) propone una modificación para un muestreo aleatorio con reemplazamiento, extrayendo una muestra de tamaño n . Para ello, este autor, realiza una pregunta indirecta para llevar a cabo un experimento aleatorio a través de esta pregunta indirecta que se presenta a continuación:

I. ¿Usted tiene la característica del grupo A (Característica sensible)?

Si su respuesta es “Sí”, sin revelar esta respuesta al entrevistador, responder con “Sí” o “No” la siguiente pregunta (II):

II. ¿Usted tiene la característica del grupo B (Característica inocua)?

De otra manera, si él/ella no tiene la característica sensible, perteneciente a la pregunta I, conteste “Sí” o “No” la siguiente pregunta (III):

III. ¿Usted tiene la característica del grupo C (Característica inocua)?

En otras palabras, en este proceso se lleva a cabo la primera cuestión y si la respuesta del encuestado, sin revelar la información acerca de esa pregunta al entrevistador, es “Sí” entonces él/ella debe contestar a la pregunta II. Si por el contrario, la respuesta es “No”, él/ella debe contestar a la pregunta III.

Claramente, los grupos con la característica inocua, ósea B y C estarán definidos por el entrevistador. Por ejemplo, el grupo B posee la pregunta acerca del último dígito del DNI y para el grupo C si el encuestado nació entre los meses de Enero y Abril.

La última modificación que ha llevado a cabo es la proporcionada por Chaudhuri y Christofides (2013), donde dichos autores han presentado un ejemplo que puede ser aplicado para estimar el porcentaje de estudiantes que han llevado un comportamiento poco ético en la universidad. Para ello, realizan dos muestras aleatorias de estudiantes en una determinada universidad. En la primera muestra de individuos los estudiantes reciben las siguientes instrucciones:

“Si durante el curso de tus estudios has pagado al menos una vez a alguien para que te haga tus trabajos asignados, contesta a la pregunta A un “Sí” o “No”. Si no es así, contesta la pregunta B.”

Pregunta A: ¿Si empezaras de nuevo tus estudios, escogerías esta Universidad para realizar sus estudios? Pregunta B: ¿Es el último dígito de su móvil 6, 7, 8 o 9?

Para la segunda muestra, los estudiantes seleccionados siguen las siguientes instrucciones:

“Si durante el curso de tus estudios has pagado al menos una vez a alguien para que te haga tus trabajos asignados, contesta a la pregunta A un “Sí” o “No”. Si no es así, contesta la pregunta B.”

Pregunta A: ¿Si empezaras de nuevo tus estudios, escogerías esta Universidad para realizar sus estudios?

Pregunta B: ¿Es el último dígito de su móvil 0, 1, 2, 3, 4 o 5?

6. ESTUDIO CON PREGUNTAS NEGATIVAS

Esta técnica fue introducida por Esponda (2006) y más tarde desarrollada por Esponda y Guerrero (2009), la cual es una técnica nueva en la investigación de aspectos sensibles, combinando elementos de investigación de estudios directos y respuesta aleatorizada. Siguiendo la propuesta de Esponda y Guerrero (2009), se considera un solo cuestionario, donde se quiere estimar un rasgo sensible a partir de únicamente una pregunta, donde los individuos participantes proporcionan una respuesta de entre las k posibles cuestiones que se presentan, aunque deben de ser excluyentes, es decir, que la respuesta que proporcione el individuo debe aplicarse con precisión a él/ella. Por ejemplo, estos autores asumieron el siguiente cuestionario:

Yo gano:

- a) Menos de 30.000 € al año.
- b) Entre 30.000 y 70.000 € al año.
- c) Más de 70.000 € al año.

Con el objetivo de estimar, con un muestra aleatoria simple de tamaño n con reemplazo, la proporción de individuos que pertenecen a cada categoría. Si se considera un cuestionario negativo al realizado:

Yo no gano:

- a) Menos de 30.000 € al año.
- b) Entre 30.000 y 70.000 € al año.
- c) Más de 70.000 € al año.

Queda claro que en este cuestionario, que dos de las tres opciones son válidas ya que, si por ejemplo, un individuo gana más de 70.000 € al año puede responder tanto (a) como (b). Por tanto, estos investigadores proponen utilizar la medida de incertidumbre de Shannon que muestra la cantidad de información revelada respondiendo a la versión directa del cuestionario.

En definitiva, se describen variaciones de estudios con preguntas negativas que pueden ser vistos en Esponda (2006) y Esponda y Guerrero (2009). Por ejemplo, en el caso donde un encuestado tiene la opción de utilizar una pregunta directa o un cuestionario con la pregunta negativa. Por otra parte, es imposible estudiar dicha técnica para estimar características puramente cualitativas.

7. USO DE TÉCNICAS INDIRECTAS EN ENCUESTAS DE SALUD

En el ámbito de la investigación en salud, las encuestas constituyen un importante instrumento de investigación, especialmente para los organismos públicos que usan este instrumento para obtener información de la población sobre los fenómenos subjetivos en relación con la salud y morbilidad percibida, las conductas relacionadas con la salud y la utilización de servicios sanitarios. Además ofrecen una visión complementaria a la obtenida con los registros de información de utilización de servicios sanitarios, que aportan una información limitada a los usuarios de dichos servicios y desde la perspectiva de los profesionales sanitarios. Casi todos los estados tienen una encuesta nacional de salud, llevada a

cabo por el organismo oficial de estadística del país, y además llevan a cabo otras encuestas suplementarias relativas a aspectos más concretos relacionados con enfermedades, hábitos saludables, nutrición, etc.

Algunas de las preguntas que se incluyen en estas encuestas pueden ser consideradas sensibles por los encuestados y como consecuencia son reacios a responder o deliberadamente dan una respuesta falsa. Así las técnicas de encuestación indirectas están siendo aplicadas para diversos estudios relacionados con la salud. A continuación se relacionan algunas de estas aplicaciones.

Las técnicas de respuesta aleatorizada han sido usadas en encuestas para estudiar el uso de fármacos, sustancias dopantes que mejoran el rendimiento, Goodstadt y Gruson (1975); Kerkvliet (1994); Simon et al. (2006); Striegel et al. (2009); Shamsipour et al. (2014). Cobo et al. (2017) usan RRT para estimar los patrones de consumo de cannabis en población universitaria.

Arnab y Singh, (2010) utilizan una variante para estudiar la medida del impacto de la infección por el VIH en Botswana. Droitcour et al. (1991) estiman la prevalencia del virus VIH en EEUU, La aplicación de estas técnicas a la estimación de la prevalencia del aborto inducido en los Estados Unidos, México, Botswana, Taiwán y Turquía ha sido estudiada en Lara et al. (2006) y en Oliveras y Letamo (2010). Los abortos entre mujeres extranjeras en Calabria han sido estudiados por Perri et al. (2015) mediante el método cruzado.

LaBrie y Earleywine (2000) y Walsh y Braithwaite (2008), usan la técnica de conteo de ítem para investigar la actividad sexual de riesgo en población homosexual. Geng et al. (2016) usan un método de respuesta aleatorizada para estimar la edad media en el primer encuentro homosexual. Krebs et al. (2011) aplicaron la técnica de conteo de ítem, para medir la prevalencia de agresión sexual. Miner (2008) explora el uso de RRT para estimar el número medio de delitos sexuales. Vakilian et al. (2014) utiliza una modificación de RRT (el modelo transversal) para estimar la prevalencia de relaciones sexuales entre los estudiantes en Irán.

REFERENCIAS

- [1] ARNAB, R. Y SINGH, S. (2010). Randomized response techniques: An application to the Botswana AIDS impact survey. **Journal of Statistical Planning and Inference**, 140, 941–953.
- [2] CHANG, H. J., WANG, C. L., Y HUANG, K. C. (2004). On estimating the proportion of a qualitative sensitive character using randomized response sampling. **Quality and Quantity**, 38, 675-680.
- [3] CHAUDHURI, A. (2000). Network and adaptive sampling with unequal probabilities. **Calcutta Statistical Association Bulletin**, 50, 237-253.
- [4] CHAUDHURI, A. (2010). Randomized response and indirect questioning techniques in surveys. **CRC Press**.
- [5] CHAUDHURI, A. Y CHRISTOFIDES, T.C. (2007). Item count technique in estimating the proportion of people with a sensitive feature. **Journal of Statistical Planning and Inference**, 137, 589-593.
- [6] CHAUDHURI, A. Y CHRISTOFIDES, T.C. (2008). Indirect questioning: how to rival randomized response techniques. **International Journal of Pure and Applied Mathematics**, 43,283-294.
- [7] CHAUDHURI, A. Y CHRISTOFIDES, T.C. (2013). **Indirect Questioning in Sample Surveys**. Springer-Verlag, Berlin, Heidelberg, Alemania.
- [8] CHAUDHURI, A., BOSE, M. Y GHOSH, J.K. (2004). An application of adaptive sampling to estimate highly localized population segments. **Journal of Statistical Planning and Inference**, 121, 175-189.
- [9] CHRISTOFIDES, T.C. (2009). Randomized response without a randomization device. **Advances and applications in Statistics**, 11, 15-28.
- [10] COBO, B., RUEDA, M. Y LÓPEZ-TORRECILLAS, F. (2016) Application of randomized response techniques for investigating cannabis use by Spanish university students. **Int J Methods Psychiatr Res**. In press.
- [11] CORSTANGE, D. (2009). Sensitive Questions, Truthful Answers? Modeling the List Experiment with LISTIT. **Political Analysis**, 17, 45-63.
- [12] COUTTS, E. Y JANN, B. (2011). Sensitive questions in online surveys: experimental results for the randomized response technique (RRT) and the unmatched count technique (UCT). **Sociological Methods and Research**, 40, 169-193.
- [13] DROITCOUR, J., CASPAR, R. A., HUBBARD, M. L., PARSLEY, T. L., VISSCHER, W. Y EZZATI, T. M. (1991). The item count technique as a method of indirect questioning: A review of its development and a case study application. **Measurement errors in surveys**, 185-210.

- [14] DROITCOUR, J.A., LARSON, E.M. Y SCHERUREN, F.J. (2001). The three card method: estimating sensitive items with permanent anonymity of response. In **Proceedings of the Social Statistics Section of the American Statistical Association**. Alexandria, VA:ASA.
- [15] ESPONDA, F. (2006). **Negative surveys**. arXiv:ST/0608176v1.
- [16] ESPONDA, F. Y GUERRERO, V.M.(2009). Survey with negative questions for sensitive items. **Statistics and Probability Letters**, 79, 2456-2461.
- [17] GJESTVANG, C. R. Y SINGH, S. (2006). A new randomized response model. **Journal of the Royal Statistical Society: Series B (Statistical Methodology)**, 68(3), 523-530.
- [18] GENG, G.Z., GAO, G., RUAN, Y.H., YU, M.R Y ZHOU, Y.H. (2016). Behavioral Risk Profile of Men Who Have Sex with Men in Beijing, China: Results from a Cross-sectional Survey with Randomized Response Techniques. **Chinese Medical Journal**, 129(5):523-529
- [19] GOODSTADT, M.S. Y GRUSON, V. (1975). The randomized response technique: a test on drug use. **JASA**, 70, 814- 818. <http://www.jstor.org/stable/2285441>
- [20] GUPTA, S., MEHTA, S., SHABBIR, J. Y DASS, B.K. (2013). Generalized scrambling in quantitative optional randomized response models. **Communications in Statistics-Theory and Methods**, 42(22), 4034-4042.
- [21] GREENBERG, B.G., ABUL-ELA, A.L.A., SIMMONS, W.R. Y HORVITZ, D.G. (1969). The unrelated question randomized response model: Theoretical framework. **Journal of the American Statistical Association**, 42(22), 520-539.
- [22] HORVITZ, D.G., SHAH, B.V. Y SIMMONS, W.R. (1967). The unrelated question RR model. **Proceedings of the Social Statistics Section of the American Statistical Association**. 65-72. Alexandria, VA: ASA.
- [23] HUSSAIN, Z. Y SHABBIR, J. (2011). On item count technique in survey sampling. **Journal of Informatics and Mathematical Sciences**, 2, 161-169.
- [24] JONH, F.A.V. St., EDWARDS-JONES, G., GIBBONS, J.M. Y JONES, J.P.G. (2010). Testing novel methods for assessing rule breaking in conservation. **Biological Conservation**, 143, 1025-1030.
- [25] KERKVLIT, J. (1994). Estimating a logit model with randomized data: The case of cocaine use. **Australian Journal of Statistics** 36, 9-20.
- [26] KIM, J. Y WARDE, W.D. (2004a). A stratified Warner's randomized response model. **Journal of Statistical Planning and Inference**, 120(1-2), 155-165.
- [27] KREBS, C.P., LINDQUIST, C.H., WARNER, T.D., FISHER, B.S., MARTIN, S.L. Y CHILDERS, J.M. (2011). Comparing Sexual Assault Prevalence Estimates Obtained With Direct and Indirect Questioning Techniques. **Violence Against Women**, 17(2) 219–235.
- [28] KUK, A.Y.C. (1990). Extensions of Mangat's randomized response model. **Journal of Statistical Planning and Inference**, 136(4),1154-1567.
- [29] KUKLINSKI, J. H., SNIDERMAN, P. M., KNIGHT, K, PIAZZA, T., TETLOCK, P. E., LAWRENCE,G. E. Y MELLERS, M. (1996). Racial Prejudice and Attitudes toward Affirmative Action. **American Journal of Political Science**, 41, 402-419.
- [30] KUTNIK, B., BELSER, P. Y DANAILOVA-TRAINOR, G. (2007). **Methodologies for global and national estimation of human trafficking victims: current and future approaches**. Working paper, 29, International Labour Office, Ginebra.
- [31] LABRIE, J. W. Y EARLEYWINE, M. (2000). Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique. **Journal of Sex Research**, 37(4), 321–326.
- [32] LARA, D., GARCÍA, S.G., ELLERTSON, C., CAMLIN, C. Y SUÁREZ, J. (2006). The measure of induced abortion in Mexico using random response technique. **Sociological Methods & Research** 35:279-301.
- [33] MANGAT, N.S. (1994). An improved randomized response strategy. **Journal of the Royal Statistical Society. Series B (Methodological)**, 93-95.
- [34] MANGAT, N.S. Y SINGH, R. (1990). An alternative randomized response procedure. **Biometrika**, 77(2).
- [35] MEHTA, S., DASS, B.K., SHABBIR, J. Y GUPTA, S.N. (2012). A three-stage optional randomized response model. **Journal of Statistical Theory and Practice**, 6(3).
- [36] MINER, M.H. (2008). Improving the measurement of criminal sexual behavior. The application of randomized response technique. **Sexual Abuse**, 20 (1): 88-101.
- [37] MILLER, J.D. (1984). **A new survey technique for studying deviant behavior**. Ph.D. Thesis, The George Washington University.
- [38] MILLER, J.D. (1985). The nominative technique: a new method of estimating heroin prevalence. **NIDA Research Monograph**, 57, 104-124

- [39] MILLER, J.D., CISIN, I.H. Y HARREL, A.V. (1986). A new technique for surveying deviant behavior: item count estimates of marijuana, cocaine and heroin. Paper presented at the annual meeting of the American Association for Public Opinion Research. **St. Petersburg, Florida.**
- [40] OLIVERAS, E. Y LETAMO, G. (2010). Examples of methods to address underreporting of induced abortion: preceding birth technique and randomized response technique. In **Methodologies for Estimating Abortion Incidence and Abortion-Related Morbidity: A Review**, Singh, S. & Remez, L. & Tartaglione, A. (Eds.), The Guttmacher Institute – International Union for the Scientific Study of the Population.
- [41] PERRI, P.F., PELLE, E. Y STRANGES, M. (2015). Estimating induced abortion and foreign irregular presence using the randomized response crossed model. **Social Indicators Research**, 1-18. Published online.
- [42] RAGHAVARAO, D. Y FEDERER, W.F. (1979). Block total response as an alternative to the randomized response method in surveys. **Journal of the Royal Statistical Society**, Series B,41, 40-45.
- [43] SHAMSIPOUR, M., YUNESIAN, M., FOTOUHI, A., JANN, B., RAHIMI-MOVAGHAR, A., ASGHARI, F. Y AKHLAGHI, A.A. (2014). Estimating the prevalence of illicit drug use among students using the crosswise model. **Substance Use & Misuse**, 49, 1303-1310.
- [44] SIMON, P., STRIEGEL, H., AUST, F., DIETZ, K. & ULRICH R. (2006). Doping in fitness sports: Estimated number of unreported cases and individual probability of doping. **Addiction**. 101, 164-1644
- [45] STRIEGEL, H., ULRICH, R. Y SIMON, P. (2010). Randomized response estimates for doping and illicit drug use in elite athletes. **Drug and Alcohol Dependence**, 106, 230–232.
- [46] THOMPSON, S.K. (1992). **Sampling**. New York: Wiley.
- [47] THOMPSON, S.K. Y SEBER, G.A.F. (1996). **Adaptive sampling**. New York: Wiley.
- [48] TAN, M.T., TIAN, G.L. Y TANG, M.L. (2009). Sample surveys with sensitive questions: a nonrandomized response approach. **American Statistician**, 63, 9-16.
- [49] TIAN, G.L., YU, J.W., TANG, M.L. Y GENG, Z. (2007). A new nonrandomized model for analyzing sensitive questions with binary outcomes. **Statistics in Medicine**, 26, 4238-4252.
- [50] TRAPPMANN, M., KRUMPAL, I., KIRCHNER, A. Y JANN, B. (2014). Item sum: A new technique for asking quantitative sensitive questions. **Journal of Survey Statistics and Methodology**, 2(1), 58-77.
- [51] VAKILIAN, K., MOUSAVI, S.V. Y KERAMAT, A. (2014). Estimation of sexual behavior in the 18-to-24-years-old Iranian youth based on a crosswise model study. **BMC Research Notes**, 7-28.
- [52] WALSH, J. A. Y BRAITHWAITE, J. (2008). Self-Reported Alcohol Consumption and Sexual Behavior in Males and Females: Using the Unmatched-Count Technique to Examine Reporting Practices of Socially Sensitive Subjects in a Sample of University Students. **Journal of Alcohol and Drug Education**, 52(2), 49–72
- [53] WARNER, S.L. (1965). Randomized Response: a survey technique for eliminating evasive answer bias. **Journal of the American Statistical Association**, 60, 63-69.
- [54] YU, J.W., TIAN, G.L. Y TANG, M.L. (2008). Two new models for survey sampling with sensitive characteristic: design and analysis. **Metrika**, 67, 251-263.

ANÁLISIS DE LA HUELLA ECOLÓGICA DEL ECUADOR: UNA COMPARACIÓN CON AMÉRICA LATINA IMPACTO EN LA BIODIVERSIDAD Y LA INCIDENCIA DEL DESARROLLO TURÍSTICO SOSTENIBLE

Jackeline Jiménez*, Ma. Paula Iñiguez*, Diana Cajamarca*,
Priscilla Massa-Sánchez** & Valentín-Alejandro Martínez-Fernández***

* Universidad Técnica Particular de Loja. Titulación de Economía.

** Universidad Técnica Particular de Loja. Departamento de Economía. pmassa@utpl.edu.ec

***Universidad de A Coruña. Departamento de Análisis Económico y Administración de Empresas.

Investigador Prometeo de la Senescyt de Ecuador con residencia en la Universidad Técnica Particular de Loja.

ABSTRACT

In recent years it has increased the interest of the international community to analyze the major environmental problems such as climate change, pollution of water, loss of biodiversity, among others. These environmental problems bring direct impact on the deterioration of the quality of human life. In this context, the ecological footprint becomes an indicator of sustainability that includes the natural resources that humans use to meet their consumption needs. This implies not only know the ecological footprint and its components, but also calculate it and consider it as an important element for decision making at individual and corporate level as well as at the level of countries and regions. In this context and using data that Global Footprint Network (2016) presents for Ecuador, Latin America and the World (2012) regarding the ecological footprint and about its components and biocapacity, this paper presents and analyzes three aspects: first, a comparison between the ecological footprint of Ecuador in 2012 with the ecological footprint of Latin America and the world; second, the ratio and weight each sub-footprint (crops, grazing, carbon, forest reserves, fishing areas and urbanized land) has on the value of national, Latin American and global ecological footprint; and third, determine the ecological surplus or deficit of Ecuador and its relationship with Latin America and the world; finally some general considerations to help guide decision making are presented. This also is particularly important in relation to specific sectors, such as the development of tourism in Ecuador, where its main asset lies in biodiversity and the need to establish policies that affect the environmental sustainability.

KEYWORDS: Ecological footprint, biocapacity, ecological surplus, ecological deficit

RESUMEN

En los últimos años se ha incrementado el interés de la comunidad internacional por analizar los grandes problemas ambientales como el cambio climático, la contaminación de los cuerpos de agua, la pérdida de biodiversidad, entre otros, que traen repercusiones directas sobre el deterioro de la calidad de la vida humana. En este contexto, la huella ecológica se convierte en un indicador de sostenibilidad que incluye los recursos naturales que el ser humano utiliza para satisfacer sus necesidades de consumo. Esto implica no sólo conocer la huella ecológica y sus componentes, sino también calcularla y considerarla como un elemento relevante para la toma de decisiones, tanto a nivel individual y empresarial como también a nivel de países y regiones. En este contexto y haciendo uso de los datos que Global Footprint Network (2016) presenta para Ecuador, América Latina y el mundo (año 2012) respecto a la huella ecológica, sus componentes y biocapacidad, en este documento se presenta y analiza tres aspectos, primero, una comparación de la huella ecológica del año 2012 del Ecuador con respecto a la huella ecológica de América Latina y el mundo; segundo, la proporción y peso que cada una de las subhuellas (cultivos, pastoreo, carbono, reservas forestales, zonas pesqueras y suelo urbanizado) tiene en el valor de la huella ecológica nacional, latinoamericana y mundial; y, tercero, superávit o déficit ecológico de Ecuador y su relación con América Latina y el mundo; finalmente se presentan algunas consideraciones generales que permitan orientar la toma de decisiones. Ello, además, cobra especial importancia en relación a determinados sectores, como es el caso del desarrollo turístico en el Ecuador, de carácter estratégico, donde su principal activo reside en la biodiversidad y la necesidad de establecer políticas que incidan en la sostenibilidad medioambiental.

1. INTRODUCCIÓN

La huella ecológica se considera el área de territorio productivo o ecosistema acuático que se requiere para satisfacer las necesidades humanas e igualmente asimilar los desechos producidos. Por su relación directa con el estilo de vida de las personas, se utiliza también para determinar la variación en el consumo y la disponibilidad de recursos en un determinado periodo de tiempo.

Para comprender los niveles tolerables de la huella ecológica, es necesario hacer referencia a la biocapacidad, considerada como la habilidad que poseen los ecosistemas para producir recursos biológicos útiles, así como para procesar los desechos que los seres humanos le devuelven. Si se compara la huella ecológica con la biocapacidad, cabe analizar la situación ecológica de un determinado territorio; y, se puede obtener como resultado un superávit si la huella ecológica es mayor a su biocapacidad, o un déficit cuando la

huella ecológica es menor a su biocapacidad. Esto permite identificar aquellas economías cuyo nivel de consumo es superior a las posibilidades que les ofrece su territorio.

La huella ecológica se define como una herramienta contable para estimar lo que una determinada población o país requiere a fin de satisfacer sus necesidades, en términos de recursos (tierra y agua) y asimilación de desechos (Wackernagel y Rees, 2001). Mediante el cálculo de este indicador se puede evaluar el impacto sobre el planeta de un determinado modo o forma de vida y al compararlo con su biocapacidad, consecuentemente permite diseñar acciones orientadas a minimizar el impacto ambiental generado por las personas.

Si para calcular la huella ecológica fuera necesario incluir todos los artículos de consumo, tipos de desechos y funciones de los ecosistemas, sería imposible su análisis debido a la enorme cantidad de información. Sin embargo, se considera una aproximación, recogiendo algunas categorías de superficie ecológicamente productivas; y así, para el cálculo de la huella ecológica, se consideran seis actividades humanas que requieren de tierra biológicamente productiva, y son mutuamente excluyentes. Estas son:

- Tierra de cultivo: comprende el área de agricultura, conformada por todos los cultivos dedicados a alimentación de personas, animales y producción.
- Tierra de pastoreo: El área de pastoreo es la cantidad de tierra que se dedica al crecimiento de ganado y elaboración de sus productos derivados. Este tipo de ecosistema es menos productivo que los demás ecosistemas, debido a su bajo potencial para producir biomasa.
- Tierras forestales: Son bosques naturales o plantaciones de las cuales se extraen la madera y sus derivados.
- Tierras de absorción de carbono: Es la cantidad de bosques o áreas de reserva destinadas para la captura de gases emitidos por la quema de combustible fósil, especialmente la captura del carbón (CO₂).
- Zona de pesca: El área de mar dedicada al desarrollo de peces y otros productos.
- Suelo urbanizado: Considera el área de construcción y asentamiento poblacional; es decir la infraestructura utilizada para la vivienda, transporte, producción industrial, etc.

A nivel mundial la huella ecológica se ha incrementado aceleradamente en los últimos años. Según datos del GFN (2016) el mundo sobrepasó su biocapacidad en el año 1970. En Ecuador constituye un indicador de sostenibilidad ambiental para estimar la manera en que los hábitos de consumo afectan al medio ambiente y ocasionan problemas como emisiones de carbono, consumo agresivo de agua, energía, comida e infraestructura, degradación del suelo, deforestación, cambio climático, pérdida de especies vertebradas, terrestres, marinas y de agua dulce.

2. EL CASO DE ECUADOR

Ecuador es uno de los países con mayor biodiversidad del mundo, y en los años 60 su biocapacidad superaba en gran medida a la demanda de recursos y, por lo tanto, la huella ecológica era mínima, con un consecuente superávit ecológico. Ahora bien, este superávit ha disminuido con el paso de los años, debido a la tendencia creciente de la huella ecológica. Además, la población se triplicó desde 1961 y la biocapacidad del país cayó a su tercera parte en 50 años (Distrito Metropolitano de Quito, 2009).

El sistema productivo ecuatoriano responde a la demanda de consumo del país y ello, a su vez, afecta directa e indirectamente al ecosistema nacional, pues la producción (en distintas ramas) implica el consumo de agua, energía, infraestructura, procesamiento de comida y contaminación con emisiones de CO₂ (Baca, Andrade, & Défaz, 2012). El sector agro-exportador ecuatoriano, se ha concentrado por décadas en la explotación de recursos naturales, afectando a la biocapacidad del país. Esto ha generado graves impactos en el territorio, tanto a nivel ambiental como social. Además, se han incrementado las actividades extractivas, ya que tan solo en 1980 y 1990 se talaron 60.000 y 340.000 hectáreas anuales, del 0,5 al 2,4% de la cobertura forestal. La contaminación por desperdicios industriales ha aumentado en los últimos años (Tomaselli, 2004). Para los años 2008 y 2009, Ecuador registró una biocapacidad de 2,35 hectáreas per cápita, y una huella ecológica de 1,62 hectáreas per cápita (Ministerio del Ambiente, 2015); si bien, para el año 2012 (últimos datos consolidados) la huella ecológica aún no sobrepasaba la biocapacidad, se evidenciaba ya una tendencia creciente de la misma. Y, a este respecto ha de significarse que de este año para acá el turismo ha experimentado un sensible crecimiento, fundamentalmente en las zonas que presentan una mayor biodiversidad, con el consiguiente impacto que, por sí mismo y dada su tendencia a un desarrollo acelerado, hace aconsejable un estudio singular de su incidencia y cómo ésta debe de armonizarse con el criterio de la sostenibilidad medioambiental.

Si bien el Ecuador, por su huella ecológica, es uno de los ocho países junto a Argelia, Colombia, Georgia, Jamaica, Jordania, Sri Lanka y Cuba, que ha demostrado la factibilidad de alcanzar el desarrollo viviendo dentro los límites del planeta (Ayala & Andes, 2015), desde 1961 la biocapacidad del país ha disminuido en términos per cápita, como consecuencia, principalmente, del aumento de la población y al desarrollo de actividades extractivas e industriales que dañan el capital natural ecuatoriano (Pérez-Neira, Marco-Larrauri & Álvarez-Muñoz, 2015). Consecuentemente, de continuar la tendencia negativa, en el futuro podría enfrentar un déficit ecológico, acentuado por la actividad en otros sectores, como el ya apuntado del turismo.

La degradación del medio ambiente, causada por los patrones de consumo, ha provocado un gran interés en la comunidad internacional. Ello se evidencia en los estudios realizados en varios países sobre huella ecológica y biocapacidad. Así, Schneider y Samaniego (2009) presentan un análisis a nivel mundial, lo cual les permite concluir que la huella de carbono representa el 50% de la HE y es el componente con un crecimiento más acelerado. El continente con mayor huella ecológica es América, con aproximadamente 9 hectáreas por persona, fundamentalmente por la degradación ambiental en América del norte.

Tomacelli (2004) determinó que la huella ecológica por persona de la Universidad de San Francisco de Quito es de 0,21 hectáreas, de lo cual el componente más influyente fue el transporte, con el consabido aumento de la huella de carbono. Por otra parte, Gómez (2009) menciona que la huella ecológica, a nivel mundial, no debería ser superior a 2.1 hectáreas por persona, que resulta de la división entre 13 mil millones de hectáreas bioproductivas disponibles para 6.5 mil millones de personas que habitamos en el planeta. No obstante, se ha registrado una huella ecológica de 2.7 hectáreas sobrepasando la biocapacidad disponible, que es de 2.1 hectáreas por persona.

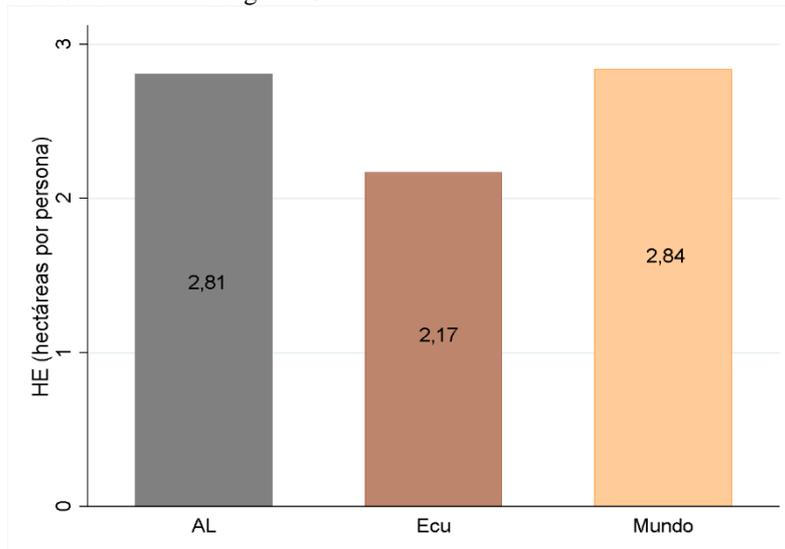
El Ministerio del Ambiente de Ecuador (2009) señala que la huella ecológica per cápita de un ecuatoriano promedio fue aproximadamente 1,6 veces menor que la huella ecológica per cápita mundial. Y según el Distrito Metropolitano de Quito (2009) el promedio de la huella de Quito es 25% más grande que el promedio dentro de Ecuador y significativamente más grande que la biocapacidad por persona disponible a nivel nacional. Si todos vivieran como un residente de Quito, necesitaríamos la biocapacidad de 1.4 planetas para soportar nuestro consumo.

En un estudio más reciente Pérez-Neira, Marco-Larrauri y Álvarez-Muñoz (2015) mencionan que, a pesar de ser Ecuador un país con una huella ecológica “media”, ésta se ha incrementado en los últimos años. Ecuador aún presenta un superávit ecológico global, al ser uno de los países más biodiversos del mundo. Sin embargo, si se mantiene la tendencia y los patrones de consumo actuales, probablemente en poco tiempo la misma sea superada. Además, el país sigue siendo exportador de biocapacidad a otros territorios, por lo cual en las últimas décadas el superávit ecológico ha disminuido de manera preocupante.

3. HUELLA ECOLÓGICA DEL ECUADOR, AMÉRICA LATINA Y DEL MUNDO

A continuación se presenta el promedio de la Huella Ecológica (HE) de Ecuador, los países de América Latina y del mundo.

Gráfico 1. Huella ecológica 2012



Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

El gráfico 1, evidencia una diferencia de 0,64 hectáreas entre la huella ecológica de Ecuador (2,17 ha/per) y la huella ecológica de América Latina (2,81 ha/per). Esto quiere decir que el nivel de consumo del Ecuador es inferior al de los países vecinos en conjunto. Por otro lado, si comparamos la huella ecológica ecuatoriana con el promedio a nivel mundial, se obtiene como resultado que la huella ecológica del mundo (2,84 ha/per) supera también a la huella ecológica de América Latina. Esto evidencia que el nivel de huella ecológica de Ecuador está por debajo del promedio de América Latina y del mundo, por lo cual su nivel de consumo per cápita afecta en menor proporción al ecosistema donde vivimos. Es necesario considerar el peso de cada componente de la huella ecológica y así determinar cuál de ellos influye más tanto en Ecuador como en América Latina y el resto del mundo. Esto se presenta en la tabla siguiente:

3.1. Componentes de la huella ecológica del Ecuador, América Latina y del mundo

Como se ha mencionado en párrafos anteriores, la huella ecológica incluye seis componentes. En la tabla 1 se presenta el valor de huella correspondiente a cada uno de ellos, tanto de Ecuador como de América latina y el mundo.

Tabla 1. Componentes de la Huella Ecológica (HE)

Componentes de HE	HUELLA ECOLÓGICA (HE)					
	Ecuador		América Latina		Mundo	
Tierras de cultivo	0,30	13,82%	0,54	19,22%	0,56	19,72%
Tierra de pastoreo	0,30	13,82%	0,57	20,28%	0,16	5,63%
Productos forestales	0,24	11,06%	0,39	13,88%	0,27	9,51%
Carbono	1,09	50,23%	1,15	40,93%	1,70	59,86%
Zona de pesca	0,18	8,29%	0,08	2,85%	0,09	3,17%
Suelo urbanizado	0,06	2,76%	0,08	2,85%	0,06	2,11%
Total HE	2,17	100%	2,81	100%	2,84	100%

Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

Para el cálculo de la huella ecológica del Ecuador corresponde el siguiente procedimiento:

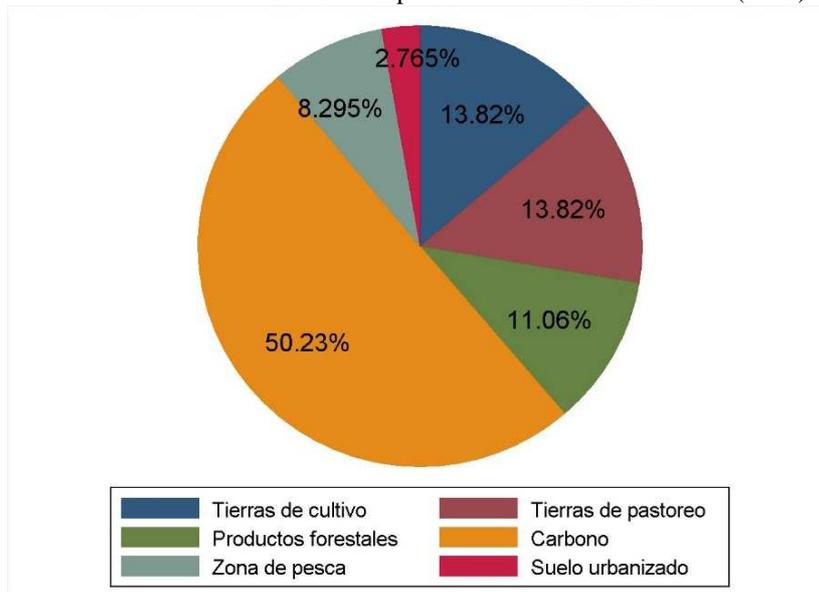
$$HE_{ECU} = cultivo_{ECU} + pastoreo_{ECU} + forestales_{ECU} + carbono_{ECU} + pesca_{ECU} + suelo.urbanizado_{ECU}$$

$$HE_{ECU} = 0,30 + 0,30 + 0,24 + 1,09 + 0,18 + 0,06$$

$$HE_{ECU} = 2,17 \text{ Ha/per}$$

El peso o proporción que tiene cada factor en la huella ecológica del Ecuador, se presenta en el siguiente gráfico:

Gráfico 2. Componentes de la HE del Ecuador (2012)



Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

En el gráfico 2 se evidencia que el factor con mayor peso en la huella ecológica ecuatoriana es la huella de carbono, al representar aproximadamente el 50%, seguido de la huella en pesquera y la huella de pastoreo representando el 13,82% cada uno; la huella de productos forestales supone el 11,06%, seguida por la huella de cultivos con el 8,29%; y por último, la huella de suelo urbanizado, con sólo el 2,76%, siendo este último el factor con menor influencia.

Se realiza la misma interpretación de los datos de los componentes de la huella ecológica de América Latina y del mundo. Por consiguiente, el valor de la huella ecológica de América Latina se ha calculado de la siguiente manera:

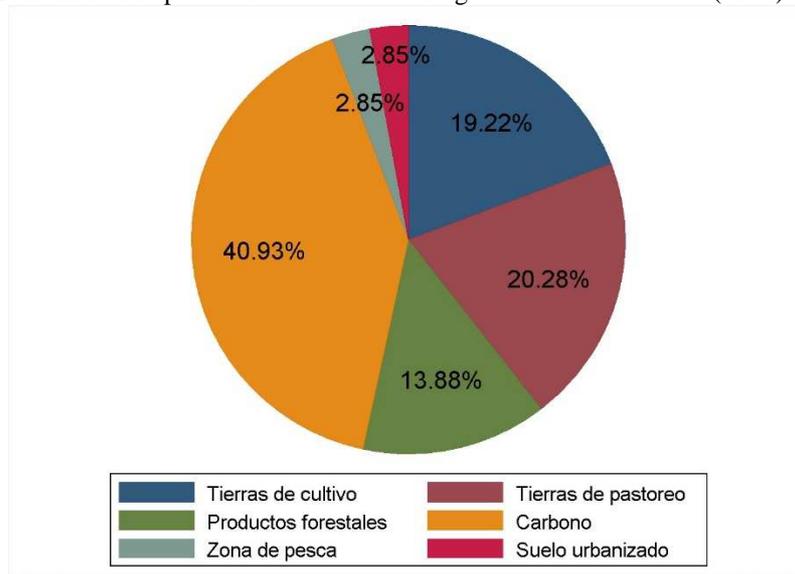
$$HE_{AL} = cultivo_{AL} + pastoreo_{AL} + forestales_{AL} + carbono_{AL} + eco.marinosa_{AL} + tierra.urbaniza_{AL}$$

$$HE_{AL} = 0,54 + 0,57 + 0,39 + 1,15 + 0,08 + 0,08$$

$$HE_{AL} = 2,81 \text{ Ha/per}$$

En el siguiente gráfico se evidencia el peso de cada factor en la HE promedio de los países de América Latina.

Gráfico 3. Componentes de la huella ecológica de América Latina (2012)



Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

El gráfico 3 muestra que el factor de más impacto en la huella ecológica de América Latina es la huella de carbono, pues supone el 40,93% de la HE, seguido de la huella de pastoreo con el 20,28%, la huella de tierras de cultivos con el 19,22% y la huella de productos esel 13,88%, mientras que la huella pesquera y de suelo urbanizado significa solamente el 2,85%, respectivamente. América Latina y Ecuador presentan características similares; es decir, el carbono es el factor que más influye y la huella de zona urbanizada el de menor intervención. Esto pone de manifiesto que tienen similares patrones de consumo. La mayoría de países latinoamericanos cuentan con una inigualable biodiversidad y amplia variedad de productos forestales y pesqueros, que generalmente son exportados a países más desarrollados; por lo tanto, el peso de estos factores en la HE no es que sean consumidos en su totalidad por sus habitantes, sino más bien son exportados a países con déficit ecológico para compensar su falta de biocapacidad y cubrir sus necesidades de recursos primarios.

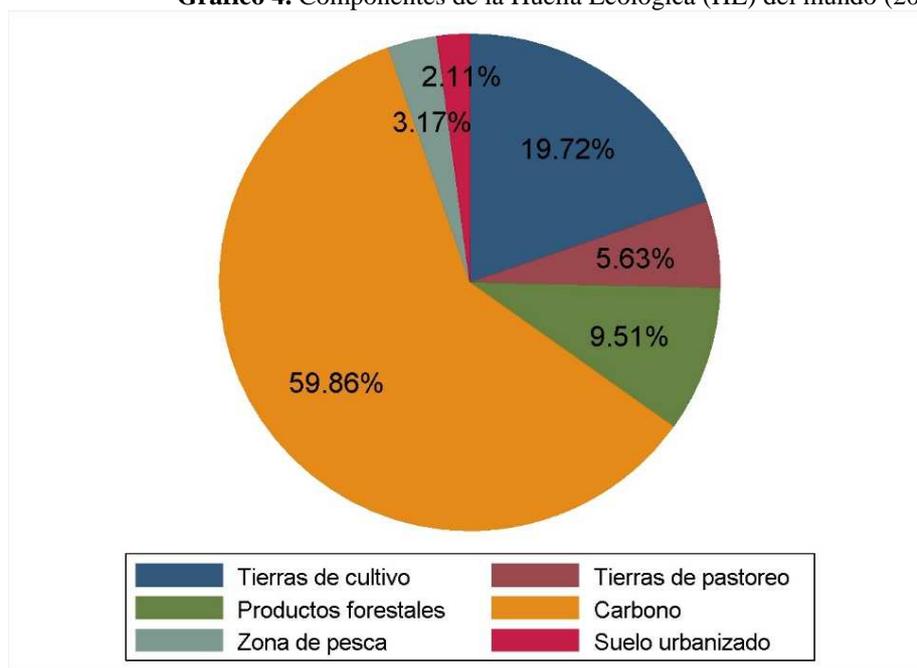
Por otro lado, el cálculo de la huella ecológica a nivel mundial, se presenta de la siguiente manera:

$$HE_{MUNDO} = cultivo_{MUNDO} + pastoreo_{MUNDO} + forestales_{MUNDO} + carbono_{MUNDO} + eco.marinosa_{MUNDO} + tierra.urbaniza_{MUNDO}$$

$$HE_{MUNDO} = 0,56 + 0,16 + 0,27 + 1,70 + 0,09 + 0,06$$

$$HE_{MUNDO} = 2,84 \text{ Ha/per}$$

El peso de cada factor de la huella ecológica se muestra a continuación:

Gráfico 4. Componentes de la Huella Ecológica (HE) del mundo (2012)

Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

Al igual que en Ecuador y América Latina, a nivel mundial el factor que más peso tiene en la huella ecológica es la huella de carbono al suponer el 59,86%. El uso especialmente de combustibles fósiles hace que las emisiones de CO₂ aumenten a un ritmo acelerado y destructivo. Esto evidencia un mundo consumista que reduce su biocapacidad y con ello las posibilidades de contar con las condiciones suficientes para que la vida pueda desarrollarse adecuadamente.

En la huella ecológica del mundo, la de cultivos representa el 19,72%, la de productos forestales con el 9,51% y con menos proporción están las de pastoreo, pesquera y de suelo urbanizado con el 5,63%, 3,17% y 2,11% respectivamente. La HE del mundo tiene a la de cultivos como el segundo factor de influencia, a diferencia de los países de América Latina cuyo segundo factor de importancia es la huella de productos forestales, debido a la riqueza forestal que poseen.

3.2. Superávit o déficit ecológico

La diferencia existente entre la biocapacidad (2,2 ha por persona) y la huella ecológica (2,17 ha por persona) refleja que el superávit ecológico en Ecuador es de 0,03 ha por persona. Si bien este país es uno de los más biodiversos, en superficie es más pequeño que otros de América Latina; de ahí que su biocapacidad sea menor en comparación con el promedio de América Latina que es de 5,3 ha por persona.

América Latina cuenta con un tercio de las reservas de agua dulce, 12% de la superficie cultivable y 21% de la superficie de bosques naturales del mundo (WWF, 2014). Sin embargo, los modelos de desarrollo basados en actividades extractivas han traído consecuencias sobre la huella ecológica que, en promedio, es de 2,8 ha por persona, mientras que la huella ecológica de Ecuador es de 0,6 ha por persona. No obstante, según el informe de Global Footprint Network, la mayoría de los países de la región reflejan superávit ecológico.

Tabla 2. Huella ecológica y biocapacidad de Ecuador, América latina y el Mundo

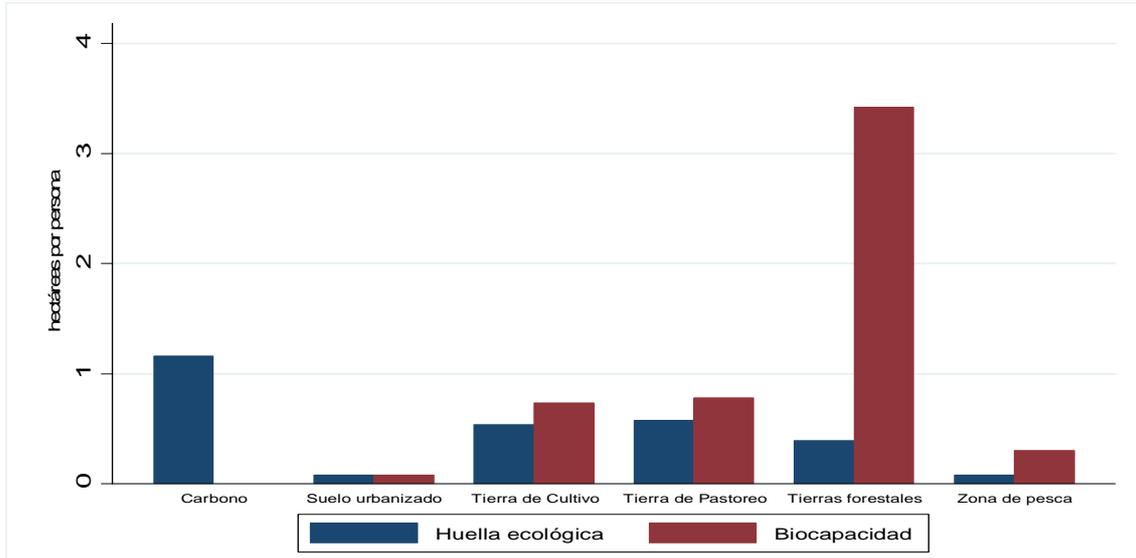
	Huella ecológica	Biocapacidad	Superávit/déficit ecológico
Ecuador	2,17	2,20	0,03
América Latina	2,81	5,31	2,50
Mundo	2,84	1,73	-1,11

Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

En la tabla 2 puede observarse que Ecuador tiene una menor huella ecológica y una mayor biocapacidad que el promedio mundial, de igual manera se puede ver que a nivel mundial existe un preocupante déficit ecológico.

Es importante señalar que las naciones más pobres y vulnerables están subsidiando el estilo de vida de los países más ricos. Así, los países de altos ingresos tienen una huella ecológica en promedio cinco veces mayor que la de países de bajos ingresos. La biodiversidad, a nivel mundial, desde 1970 ha disminuido más rápido en los países de bajos ingresos (WWF, 2014). Así, la disminución de la biocapacidad hace necesario que un país importe recursos esenciales de los ecosistemas extranjeros.

Gráfico 5. Huella Ecológica para Ecuador y sus componentes

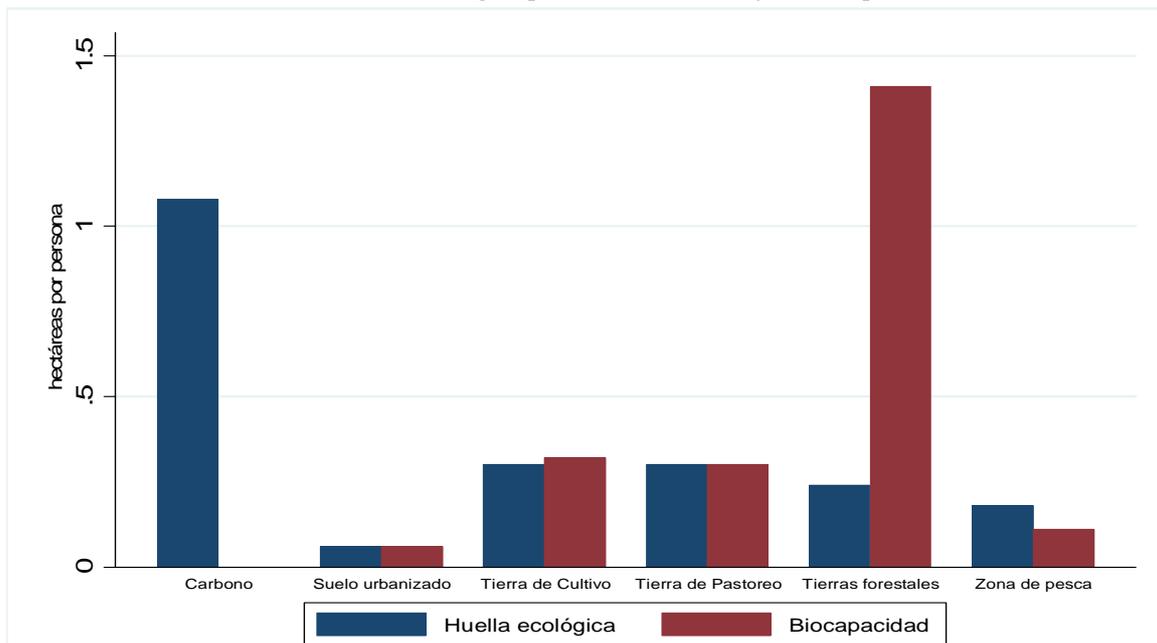


Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

El aumento de la demanda de recursos naturales se ha vuelto insostenible, como se ve reflejado en la huella ecológica. Al comparar los componentes de la huella ecológica con su biocapacidad (ver Gráfico 5), se observa que existe un superávit ecológico en tierra de cultivo y tierras forestales. Mientras que la tierra de pastoreo y suelo urbanizado presenta el mismo nivel entre la HE y su biocapacidad. En cuanto a la zona de pesca, refleja un déficit ecológico de 0,07 ha per cápita. Esta situación se da porque la actividad económica del país se basa en el sector primario.

También se observa el impacto de la urbanización con sus correspondientes problemas ambientales, relacionados con la disminución de cobertura vegetal, incremento en el consumo de recursos como agua, energía, acumulación de basura y contaminación industrial. Por otra parte, las tierras forestales muestran un superávit de 1,17 has per cápita, al existir importantes áreas forestales ubicadas en distintos lugares del Ecuador.

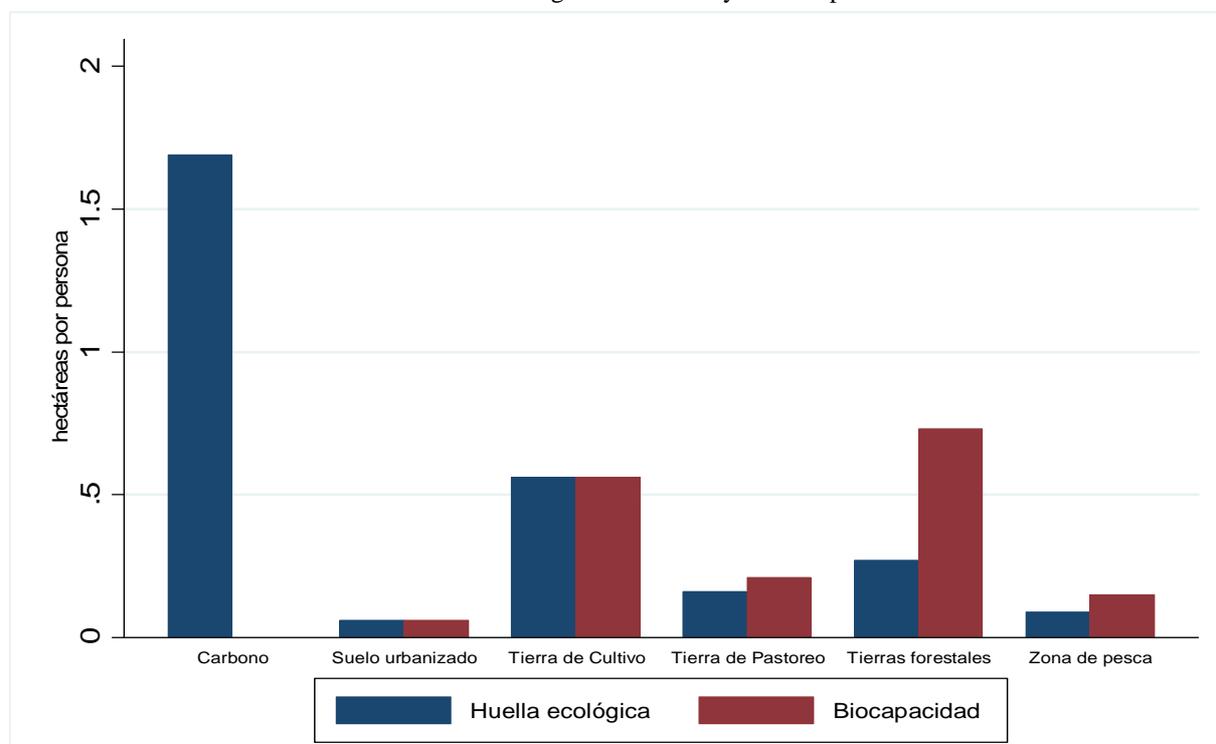
Gráfica 6. Huella Ecológica para América Latina y sus componentes



Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

En el Gráfico 6, se evidencia la presencia de un superávit ecológico en tierra de cultivo, de pastoreo, tierras forestales y zona de pesca. Y de la misma manera, el suelo urbanizado presenta el mismo nivel entre la huella ecológica y su biocapacidad. En cuanto a la tierra de cultivo, aunque la mayoría de países de América Latina basan sus economías en actividades extractivistas, se tiene tan solo un 0,19 has per cápita de capacidad para generar materiales orgánicos y captar los desechos que ocasionan las actividades humanas.

Gráfico 7. Huella Ecológica del mundo y sus componentes



Fuente: Elaboración propia, a partir de Global Footprint Network (2016)

En la gráfica 7 se observa que la biocapacidad y la huella ecológica de las tierras de cultivo y del suelo urbanizado se encuentran al mismo nivel. Mientras que la tierra de pastoreo, tierra forestal y zona pesquera cuentan con superávit ecológico. La huella de carbono, a nivel mundial supera a la huella de Ecuador, como de la de América Latina.

4. CONSIDERACIONES GENERALES

La huella ecológica produce impactos negativos en el medio ambiente y el consumo acelerado de recursos naturales afecta la biocapacidad de Ecuador, América Latina y el mundo.

En Ecuador, la huella ecológica ha aumentado con el paso de los años mientras que su biocapacidad ha disminuido. Esto como consecuencia del modelo de desarrollo agro-exportador concentrado por décadas en la explotación de recursos naturales. Si bien, estos recursos se orientan al consumo interno, gran parte de ellos se destinan al consumo de países con déficit ecológico, convirtiendo así al Ecuador en un país exportador de biocapacidad.

Es necesario desarrollar procesos de concienciación a la población, sobre hábitos de consumo ecológicamente responsables, tales como: uso de energías limpias, ahorro de energía, consumo responsable del agua, manejo adecuado de residuos y reciclaje, entre otros; y, de esta manera, frenar la disminución paulatina de biocapacidad en el Ecuador, y así contribuir a que mantenga el superávit ecológico.

Para mitigar los efectos agravantes que ocasionan los patrones de consumo en Ecuador, es necesaria la implementación de políticas encaminadas al correcto uso de los recursos naturales, enfocadas en la evaluación, minimización y mitigación de sus impactos tanto ambientales, como sociales y económicos, esto con la finalidad de usar de manera sostenible los recursos naturales. En este sentido, resulta de especial importancia la definición de políticas públicas orientadas a diseñar planes de actuación de desarrollo turístico basados en la preservación de la biodiversidad, de modo que se minimice el impacto que pudiera tener en el medio ambiente la creación de infraestructuras dirigidas a la prestación de servicios turísticos. El énfasis en este sector se justifica en el hecho de que tiene un carácter de tipo estratégico y todo indica que, en los próximos años, conforme al crecimiento acelerado de los últimos, se situará entre las tres principales fuentes de ingresos del país.

Ante esta situación, la vía más eficaz para disminuir los impactos que provoca el consumo masivo y una elevada huella ecológica, es imprescindible la colaboración entre entidades públicas, privadas y organizaciones no gubernamentales; así como también, la generación de una cultura cívica, respecto a los patrones de consumo y contaminación del ambiente.

Mención de Reconocimiento: La contribución del Ph.D. Valentín Alejandro Martínez Fernández al presente trabajo se enmarca dentro de su actividad como investigador del Programa Prometeo, financiado por la Secretaría. Dentro de sus actividades desarrolladas en el área de Administración de la Universidad Técnica Particular de Loja.

REFERENCIAS

- [1] Andes (2015). Ecuador, los efectos del cambio climático, iniciativas ambientales y la COP 21. Andes, págs. 1,2.
- [2] Baca, J. C., Andrade, A. K., & Défaz, S. (2012). Informe de Cálculo de la Huella Ecológica del Ministerio del Ambiente. Disponible (on line): <http://www.ambiente.gob.ec/wp-content/uploads/downloads/2013/04/Informe-Huella-Ecol%C3%81gica-Ministerio-del-Ambiente-2012-Final...pdf>
- [3] Bastida, E. L. (2010). La huella ecológica como indicador de sostenibilidad. Disponible (on line): <http://www.cubasolar.cu/biblioteca/Energia/Energia54/HTML/Articulo09.html>
- [4] Distrito Metropolitano de Quito. (2009). Huella Ecológica de Quito. Disponible (on line): http://www.footprintnetwork.org/images/article_uploads/Huella_Ecol%C3%B3gica_Quito_imp_28_sep_11.pdf
- [5] Global Footprint Network. (2016). Disponible (on line): http://www.footprintnetwork.org/en/index.php/GFN/page/case_stories/#qui
- [6] Goldfinger, Wackernagel, Stechbart, Rizk, & Kitzes., R. &. (2008). The Ecological Footprint. Disponible (on line): <http://www.footprintnetwork.org/es/index.php/GFN/page/trends/ecuador/>
- [7] Gómez, D. (2009). La huella ecológica y los países andinos, una reflexión sobre la sustentabilidad. Obtenido de http://www.flacsoandes.edu.ec/sites/default/files/agora/files/1260980537.huella_ecologica_deyanira_gomez_0.pdf
- [8] Ibáñez. (s.f.). Huella Ecológica. Obtenido de http://www.academia.edu/5204302/HUELLA_ECOLOGICA
- [9] López, E. (s.f). ¿Cuántos recursos naturales utilizan las personas para mantener su estilo de vida?. Obtenido de <http://www.cubasolar.cu/biblioteca/Energia/Energia54/HTML/Articulo09.html>
- [10] Ministerio del Ambiente del Ecuador (2013). “Reporte de la Huella Ecológica del Ecuador: 2008 y 2009”. Primera edición, Quito - Ecuador.
- [11] Moore, D. (2009). Huella ecológica de Quito. Disponible (on line): http://www.footprintnetwork.org/images/article_uploads/Huella_Ecol%C3%B3gica_Quito_imp_28_sep_11.pdf
- [12] Neira, & Muñoz, L. &. (2015). La huella ecológica de las naciones. Obtenido de <http://www.unemi.edu.ec/ojs/index.php/cienciaunemi/article/viewFile/184/1>
- [13] Schneider, H., & Samaniego, J. (2010). La huella del carbono en la producción, distribución y consumo de bienes y servicios. Santiago de Chile: Comisión Económica para América Latina y el Caribe (CEPAL), Naciones Unidas. Disponible (on line): http://repositorio.cepal.org/bitstream/handle/11362/3753/1/S2009834_es.pdf
- [14] Tomaselli Crespo, M. F. (2004). Investigación de la huella ecológica en la Universidad San Francisco: cálculo y creación de un reportaje.
- [15] Wackernagel, M., & Rees, W. E. (2001). Nuestra huella ecológica: Reduciendo el impacto humano sobre la Tierra. Lom Ediciones.
- [16] WWF. Informe Planeta Vivo 2014. Disponible (on line): http://d2ouvy59p0dg6k.cloudfront.net/downloads/informe_planetavivo2014_lowres2_1.pdf

ESTIMACION DE LA SUPERVIVENCIA EN PACIENTES CON CANCER DE PULMON TRATADOS CON VACUNAS TERAPEUTICAS CUBANAS.

Carmen Viada-González*, Martha Fors-López**, Patricia Lorenzo-Luaces*,
Lizet Sánchez Valdés*, Mabel Alvarez-Cardona*, Aliuska Frías-Blanco*,
Lázara García-Fernández*, Yanela Santiesteban-González* & Yuliannis Santiesteban-
González*.

1. Centro de Inmunología Molecular, Habana, Cuba
2. Escuela de Medicina, Universidad de las Américas, Quito, Ecuador

ABSTRACT:

Introduction: CIMAvaxEGF and Cuban Racotumomab are therapeutic vaccines licensed for use in adult patients with lung cancer non-small cell (NSCLC) in stage IIIB / IV. Objective: To analyze the predictability of survival in patients with this diagnosis. Materials and methods: A search strategy for clinical studies published in peer-reviewed international scientific journals has been implemented. 17 articles in PubMed between January 1, 1994 and September 30, 2015, including 12 clinical trials in Cuba from 1995 to 2015 with CIMAvaxEGF that included 1981 patients and six clinical trials were obtained Racotumomab, which included 521 subjects a methodology decision para tree predictability analysis of survival data mining technique was used. Results: Patients with lung cancer treated with the vaccine had been CIMAvaxEGF generally less than or equal to 2 and female had survival of 15.88 months. Patients with lung cancer treated with the vaccine Racotumomab had higher survival (14.39 months) were those who had just 0-1 overall and stage IIIB. Conclusions: The construction of decision trees is effective for classification of subject and for predicting certain variables. In this case specifically, it is an effective tool for predicting chances of survival in patients with lung cancer treated with Cuban therapeutic vaccines

KEYWORDS: Data mining, decision trees, epidermal growth factor, non-small cell lung cancer, therapeutic vaccine

RESUMEN:

Introducción: CIMAvaxEGF y Racotumomab son vacunas terapéuticas cubanas licenciadas para su uso en pacientes adultos con cáncer de pulmón de células no pequeñas (CPCNP) en etapas IIIB/IV. Objetivo: Analizar la predictibilidad de supervivencia en pacientes con este diagnóstico. Materiales y métodos: Se implementó una estrategia de búsqueda de estudios clínicos publicados en revistas científicas internacionales arbitradas. Se obtuvieron 17 artículos en Pubmed entre el 1 de enero de 1994 y 30 de septiembre del 2015, de ellos 12 ensayos clínicos realizados en Cuba de 1995-2015 con CIMAvaxEGF que incluyeron 1981 pacientes y seis ensayos clínicos con Racotumomab, los cuales incluyeron 521 sujetos, se utilizó una metodología de árbol de decisión para el análisis de la predictibilidad de la supervivencia, técnica de minería de datos. Resultados: Los pacientes con cáncer de pulmón tratados con la vacuna CIMAvaxEGF que tenían estado general menor o igual que 2 y sexo femenino tuvieron supervivencia de 15.88 meses. Los pacientes con cáncer de pulmón tratados con la vacuna Racotumomab que tuvieron mayor supervivencia (14.39 meses) fueron los que tenían estado general igual 0-1 y estadio IIIB. Conclusiones: La técnica de la construcción de árboles de decisión es efectiva para la clasificación de sujetos y para la predicción de ciertas variables. En este caso específicamente, es una herramienta efectiva para la predicción de probabilidades de supervivencia en pacientes con cáncer de pulmón tratados con vacunas terapéuticas cubanas

1. INTRODUCCIÓN

La minería de datos se puede considerar como un proceso para seleccionar, explorar y modelar gran cantidad de datos y descubrir patrones desconocidos, aspecto extremadamente importante en el caso de los ensayos clínicos de nuevos medicamentos. Aunque en la industria farmacéutica los datos generados no son tan grandes como en otras industrias, existen numerosos datos que pueden ser analizados a través de técnicas minería de datos. Específicamente, la minería de datos es un proceso computarizado para la utilización de la gran cantidad de conocimiento existente en la literatura o en diferentes textos que se revisan, analizando información de alta calidad, la cual se puede obtener a través de la elaboración de patrones y tendencias estadísticas.

El tiempo de predicción de la supervivencia en el momento del diagnóstico del cáncer es de gran importancia para tomar decisiones sobre el tratamiento y la atención de seguimiento a largo plazo. Sin em-

bargo, la predicción del resultado de cáncer sobre la base de la información clínica es una tarea difícil, por lo que se buscan diversas alternativas tanto estadísticas como informáticas, entre las cuales están los diferentes algoritmos de la minería de datos para predecir la supervivencia de pacientes con cáncer de pulmón que recibieron dos diferentes vacunas terapéuticas cubanas

En Cuba como en Ecuador, el cáncer es la segunda causa principal de muerte y la primera causa de años de vida potencialmente perdidos, lo cual ocasiona un impacto significativo en la esperanza de vida al nacer. El cáncer de pulmón es la enfermedad maligna de mayor incidencia y además la causa principal de muerte por cáncer en el país. En los años 2005–2007, en Cuba, se reportó anualmente un promedio de 4234 nuevos casos de cáncer de pulmón y 4601 muertes por la enfermedad, para una tasa cruda de mortalidad de 54,3 en hombres y 27,3 en mujeres por 100 000 habitantes [Galan, 2009].

Según GLOBOCAN 2012, un proyecto de la Organización Mundial de la Salud, se estimó que, en ese año, en Ecuador la incidencia de cáncer en hombres fue de 162 por 100.000 habitantes, por encima de países como Perú (140,9) o Venezuela (146,9), pero por debajo de Brasil (231,6) o Cuba (250,8). [Romero, 2009]

Si persisten las tendencias demográficas actuales y los factores de riesgo, se puede esperar que la incidencia y la mortalidad por cáncer de pulmón aumenten significativamente en los próximos cinco años, en ambos países.

Considerando que el aumento de la incidencia y la mortalidad por cáncer es un serio problema para la salud pública, el Ministerio de Salud Pública de Cuba (MINSAP) ha implementado un Programa Integral de Control del Cáncer (PICC), que opera a todos los niveles del sistema nacional de salud. Este programa constituye un nuevo enfoque terapéutico para la enfermedad, donde la biotecnología sirve de puente entre la investigación en inmunología básica y la salud pública. En Cuba la investigación y el desarrollo de la biotecnología se conciben como un ciclo científico completo (ciclo cerrado), desde el concepto hasta la aplicación clínica y comercialización de los productos. Los ingresos por las ventas del producto se invierten tanto en sostener el uso de las inmunoterapias en el sistema nacional de salud pública, como en la investigación y desarrollo de nuevas modalidades inmunoterapéuticas [Lage 2008 y 2009].

Por su parte el Ministerio de Salud Pública del Ecuador (MSP) conformó el Comité Ecuatoriano Multisectorial del Cáncer (CEMC); instancia de gestión que integra varias instituciones, públicas, privadas y de la sociedad civil en un equipo interinstitucional que permitirá la armonización y gobernabilidad, de la lucha contra el cáncer, bajo la rectoría del Ministerio de Salud. El MSP brinda los servicios oncológicos en los hospitales Eugenio Espejo (Quito), Baca Ortiz (Quito), Francisco de Ycaza Bustamante (Guayaquil) y adicionalmente se realizó la apertura en el 2013 de dos centros de atención integral y acogida para personas con cáncer en Quito y Cuenca.[Gerard, 2009]

Aunque las quimioterapias se han convertido en un arsenal indispensable para la reducción de la carga tumoral y el aumento de la supervivencia, su impacto en el cáncer de pulmón solo es medible en meses, con graves reacciones adversas como carga adicional. El mecanismo de acción de los productos inmunoterapéuticos, tales como los anticuerpos monoclonales (AcM) y las vacunas terapéuticas, es mucho más selectivo en la eliminación de las células tumorales y puede aumentar la supervivencia de los pacientes proporcionándoles una mejor calidad de vida.[Lage 2009, Toffoli 2007, Hirsch 2009]

El receptor del factor de crecimiento epidérmico (REGF) es un oncogén bien conocido. La vía de señalización y transducción del REGF puede ser interrumpida eficazmente por la privación de EGF, inhibición directa del receptor con AcM específicos, o moléculas de bajo peso molecular que compiten intracelularmente con el adenosín trifosfato (ATP) por el sitio de actividad tirosina quinasa del receptor, con repercusiones negativas sobre la proliferación celular y, por consiguiente, sobre el desarrollo del tumor.[González 2007, Macías 1987]

La inducción de la privación de EGF por la inmunoterapia activa es un concepto emergente desarrollado por investigadores cubanos, que incluye la manipulación de la respuesta inmune del individuo para que genere sus propios anticuerpos efectoros (Acs) contra el EGF, reduciendo de esta forma el tamaño del tumor o impidiendo su progresión.[Arteaga 2003, Lage 2003, Pérez 1984, González 1996, 1997 y 1998]

CIMAvax EGF es una vacuna terapéutica contra el cáncer desarrollado totalmente en Cuba. Sus ingredientes farmacéuticos activos (IFA) se producen en el Centro de Ingeniería Genética y Biotecnología (CIGB). La vacuna se formula en el Centro de Inmunología Molecular (CIM) y los ensayos clínicos se realizan en hospitales que reúnen las condiciones profesionales y tecnológicas para cumplir las normas de Buenas Prácticas Clínicas [Hirsch 2008, González 2003]. Desde 1995, CIMAvax se ha sometido a cinco ensayos clínicos fase I/II y uno fase II. Los resultados de estas investigaciones condujeron a que el Centro para el Control Estatal de la Calidad de los Medicamentos (CECMED), autoridad reguladora cubana, concediera la licencia a esta vacuna terapéutica para su uso en pacientes adultos con cáncer de pulmón de células no pe-

queñas (CPCNP) en estadios IIIB/IV. Los resultados de las investigaciones también se han publicado en varias revistas científicas nacionales e internacionales.

El objetivo de este artículo es describir la supervivencia de estos pacientes a partir de un algoritmo para la construcción de un árbol de decisión, para esto se formó un equipo constituido por especialistas en estadística e informática de Cuba (Centro de Inmunología Molecular) y Ecuador (Universidad de las Américas).

2. MATERIALES Y MÉTODOS

El método usado en este trabajo fue en primer lugar la identificación de diferentes bases de datos para la aplicación del algoritmo de un árbol de decisión, el cual es un método de extracción de datos de uso común para el establecimiento de sistemas de clasificación basados en múltiples covariables. Un árbol de decisión consiste en nodos que describen todas las alternativas posibles y donde se comparan todos los posibles desenlaces esperables para cada una de ellas. En este estudio, se realizó una búsqueda automatizada utilizando las palabras clave “vacuna”, “CIMvaxEGF” y “Racotumomab”, en la base de datos PubMed (Medline) a través del servicio HINARI de la OMS. Como criterios generales para la selección de los artículos se asumió que fueran publicados en revistas internacionales arbitradas. Se encontraron 17 artículos publicados entre el 1 de enero de 1994 y el 30 de septiembre del 2009.

Los árboles de clasificación fueron construidos con el paquete estadístico IBM® SPSS® Statistics Versión 21 y el método de crecimiento CHAID, que se caracteriza por detección automática de interacciones. Detección automática de interacciones mediante chi-cuadrado (CHI-square Automatic Interaction Detection). En cada paso, CHAID elige la variable independiente (predictora) que presenta la interacción más fuerte con la variable dependiente. Las categorías de cada predictor se funden si no son significativamente distintas respecto a la variable dependiente.

La información sobre variables demográficas y relativas al estado del paciente así como relativas al tumor y a la respuesta al tratamiento se utilizó para la construcción del árbol de decisión. Se conformaron dos bases de datos, una para cada una de las vacunas estudiadas, por lo que en este estudio analizaremos dos árboles de clasificación, uno por cada vacuna.

Una de las bases de datos estuvo conformada por los datos obtenidos durante la ejecución de 12 estudios de diferentes fases, se incluyeron 1981 pacientes con cáncer de pulmón los cuales recibieron CIMAvaxEGF; cinco ensayos Fase I, un Fase II, dos estudios Fase III pacientes, un Fase IV con, un estudio observacional y dos estudios promovidos por el investigador (Tabla 1).

Tabla 1. Ensayos clínicos identificados con la vacuna terapéutica anticancerosa de EGF, 1995–2015

Ensayos	N	Tratamiento	Mediana de Supervivencia (meses)	Fuente
EC019 Piloto I	10	5 EGF /TT/AL	3.0	[González, 2003]
		5 EGF/P64k/AL Esquema QVV	3.0	
EC025 Piloto II	34	17 EGF/P64k/AL	4.1	[González, 2003]
		17 EGF/P64k/Montanide ISA 51 Esquema QVV	5.4	
EC033 Piloto III	25	13 EGF/P64k/AL	4.3	[González, 2003]
		12 EGF/P64k/Montanide ISA 51 Esquema QVV	7.1	
EC041 Piloto IV	43	22 EGF/P64k/AL (1 deltoide) vs	6.4	[Crombet, 2006]
		21 EGF/P64k/AL (2 deltoides) Esquema QVV	7.1	
EC056 Fase II	80	40 EGF/P64k/Montanide ISA 51	12.7	[Neninger 2008, García 2008]
		40 Control (4 sitios) Esquema QVV	8.5	
EC062 Piloto V	21	EGF/P64k/Montanide ISA 51 (4 sitios: (2 deltoides, 2 glúteos) Esquema VQV	12.8	[Neninger, 2009]
EC081 Fase III	405	270EGF/P64k/Montanide ISA	10.4	[Viada 2013, Rodríguez 2016]
		51 135 Control	8.9	

EC111 Fase III	116	(4 sitios) Esquema QVV 58 EGF/P64k/Montanide ISA 51 58 Control (4 sitios) Esquema VQV	11.7 10.4	No Publicado
EC120 Fase IV	108 4	EGF/P64k/Montanide ISA 51 (4 sitios) Esquema QVV	7.8	[del Cristo 2015, Hernández 2016]
PL004 IIIA	21	21EGF/P64k/Montanide ISA 51 (4 sitios) Esquema QVV	8.6	No Publicado
PL008 IIIA	14	14 EGF/P64k/Montanide ISA 51 (4 sitios) Esquema QVV	8.9	No Publicado
PUCE107	128	128 EGF/P64k/Montanide ISA 51 (4 sitios) Esquema QVV	6.6	No Publicado

La otra base de datos estuvo conformada por los ensayos clínicos que evaluaron Racotumomab, donde se incluyeron 582 pacientes provenientes de seis ensayos clínicos; un Fase I, un Fase I-II, un Fase II-III, un Fase III, un estudio observacional y un estudio promovido por el investigador (Tabla 2).

Tabla 2. Ensayos clínicos identificados con la vacuna terapéutica anticancerosa de 1E10, 1995–2015

Ensayos	N	Tratamiento	Mediana de Supervivencia (meses)	Fuente
Fase I	71	71 1E10 Esquema QVV	9.93	[Alfonso, 2007]
EC080 Fase II-III	17 6	87 1E10 89 Control Esquema QVV	8.23 6.80	[Alfonso, 2009]
EC133 Fase I/II	20	20 EGF/P64k/Montanide ISA51+1E10 Esquema QVV	6.7	[Mazorra, 2014]
EC147 Fase III	96	96 1E10 Esquema QVV	9.2	No publicado
PL010	19	10 1E10 9 Control Esquema VQV	13.4 9.8	No publicado
PUCE106	20 0	EGF/P64k/Montanide ISA 51 (4 sitios: (2 deltoideos, 2 glúteos) Esquema VQV	8.96	[Santesteban 2014, Pérez 2013]

3. RESULTADOS

Para la vacuna CIMAvaxEGF, se generó un árbol de clasificación tomando como variable dependiente: la supervivencia desde la inclusión y como variables independientes las variables pronóstico: Edad, Sexo, Raza, Estadío, Estado (OMS), Hábito tabáquico, Consumo de alcohol y Localización del Tumor. (Figura 1)

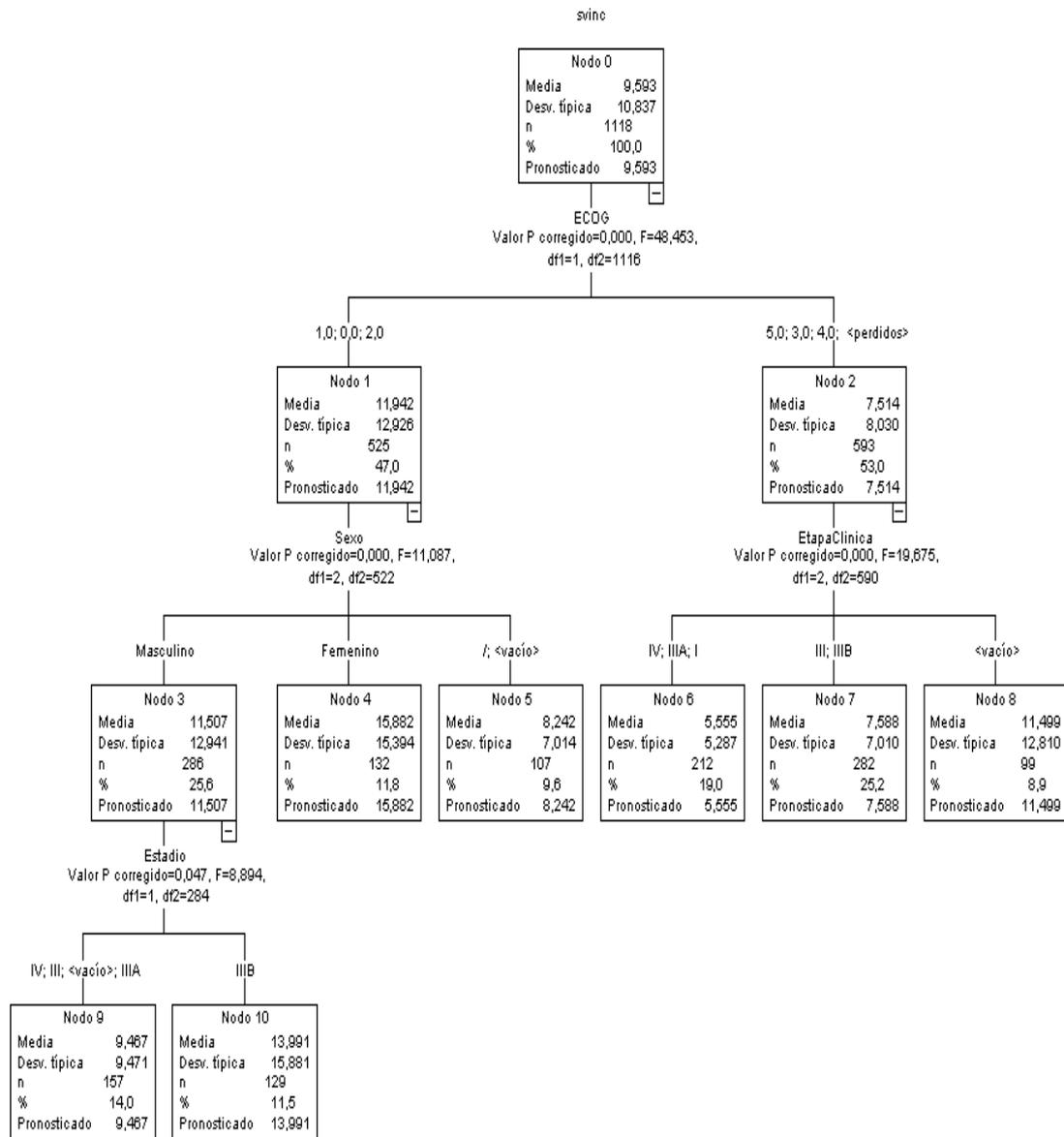


Figura. 1: Árbol de Clasificación para los pacientes con cáncer de pulmón tratado con CIMAvaxEGF

Los resultados de un ensayo clínico en oncología se expresan en términos de supervivencia o supervivencia. Con la administración de esta vacuna, se observa que la media de la supervivencia de los pacientes con estado general menor o igual que 2, sexo masculino y estadio IIIA o IV fue 9.47 meses, mientras que para los pacientes con estado general menor o igual que 2, sexo masculino y estadio IIIB, la media de la supervivencia de los pacientes fue de 13.99 meses.

La media de la supervivencia de los pacientes con estado general menor o igual que 2, sexo femenino fue de 15.88 meses y la media de la supervivencia de los pacientes con Estado general mayor que 3 y Etapa clínica IIIA o IV fue 5.55 meses. La media de la supervivencia de los pacientes con Estado general mayor que 3 y Etapa clínica III o IIIB fue 7.58 meses.

Para el caso de la administración de Racotumomab, se construyó un árbol de clasificación tomando como variable dependiente: la supervivencia desde la inclusión y como variables independientes las variables pronóstico: Edad, Sexo, Raza, Índice de Karnofsky, Tipo Histológico, Diagnóstico de la Enfermedad y Esquema de Tratamiento. (Figura 2)

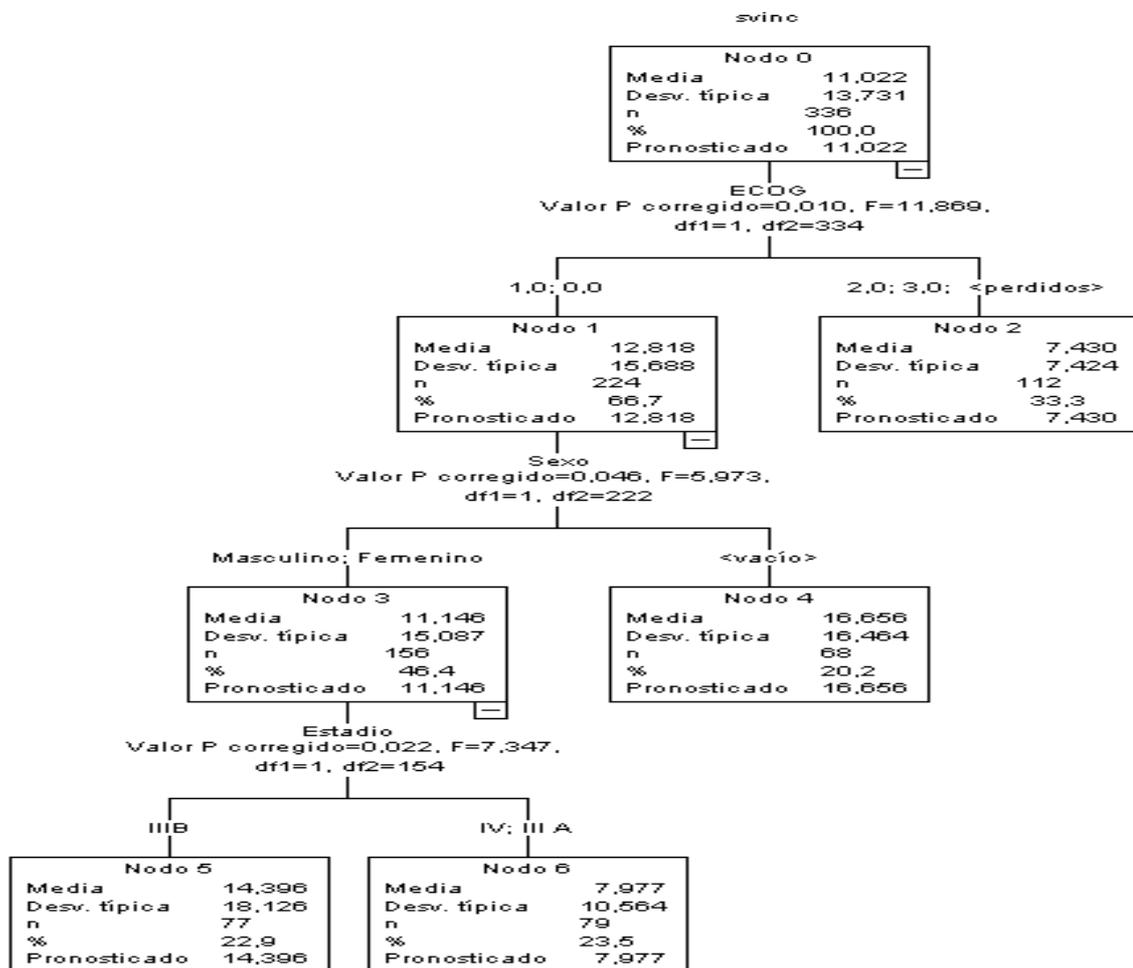


Figura. 2: Árbol de Clasificación para los pacientes con cáncer de pulmón tratado con Racotumomab

Como se observa la media de la supervivencia de los pacientes con Estado (OMS) igual 0 ó 1 y Estado IIIB fue 14.39 meses. La media de la supervivencia de los pacientes con Estado (OMS) igual 0 ó 1 y Estado IIIA o IV fue de 7.97 meses. La media de la supervivencia de los pacientes con Estado (OMS) igual 2 ó 3 fue de 7.43 meses.

3. DISCUSIÓN

El análisis particional recursivo es método no paramétrico bien establecido y su aplicación en el campo de la oncología data de más de 20 años, sin embargo, aún se utiliza para enfrentar diferentes aspectos de los ensayos clínicos de cáncer, sobre todo para identificar grupos de pacientes según factores pronóstico y para realizar predicciones sobre todo de las variables relacionadas con la supervivencia de estos pacientes. Una ventaja particular del método, en relación a otras técnicas estadísticas como el Análisis de Discriminante y la Regresión logística, es la fácil interpretación de sus resultados. Con ambas vacunas se obtuvieron cifras de supervivencias mayores a 14 meses, los pacientes con cáncer de pulmón tratados con la vacuna CIMAvaxEGF que tenían estado general menor o igual que 2 y sexo femenino tuvieron mayor supervivencia de 15.88 meses y los pacientes tratados con la vacuna Racotumomab que tuvieron mayor supervivencia (14.39 meses) fueron aquellos con Estado (OMS) igual 0 ó 1 y Estado de la enfermedad IIIB.

3. CONCLUSIONES

En la minería de datos, la técnica de la construcción de árboles de decisión es efectiva para la clasificación de sujetos y para la predicción de ciertas variables. En este caso específicamente, es una herramienta efectiva para la predicción de probabilidades de supervivencia en pacientes con cáncer de pulmón tratados con vacunas terapéuticas cubanas.

AGRADECIMIENTOS: LOS autores quieren expresar sus reconocimientos y agradecimientos en relación con el trabajo que se presenta a:

Promotores: Tania Crombet Ramos, Mayra Ramos Suzarte, Amparo Macías Abraham, Giselle Saurez, Patricia Piedra, Patricia Sierra y Normando Iznaga

Estadísticos: Patricia Lorenzo-Luaces, Mayté Robaina García, Rolando Uranga Piña, Lizet Sánchez Valdés.

Monitores: Liana Martínez Pérez, Bárbara Wilkinson Brito, Mayelin Troche Concepción y Maylen Arencibia.

REFERENCIAS

- [1] Alfonso S, Diaz RM, de la Torre A, Santiesteban E, Aguirre F, Perez K, et al. 1e10 anti-idiotypic vaccine in non-small cell lung cancer: Experience in stage iiib/iv patients. *Cancer Biology & Therapy* 2007;6(12):1847-1852.
- [2] Alfonso S, Valdés A, Santiesteban ER, Flores YI, Areces F, Hernández M, et al. A randomized, multicenter, placebo-controlled clinical trial of racotumomab-alum vaccine as switch maintenance therapy in advanced non-small-cell-lung cancer patients. *Clinical Cancer Research*, 2014 Jul 15; 20(14): 3660-71.
- [3] Arteaga CL. ErbB-targeted therapeutic approaches in human cancer. *Exp Cell Res*. 2003 Mar 10;284(1):122-30.
- [4] Aucouturier J, Ascarateil S, Dupuis L. The use of oil adjuvants in therapeutic vaccines. *Vaccine*. 2006 Apr 12;24 Suppl 2:S2-44-5.
- [5] Berd D, Sato T, Maguire HC Jr, Kairys J, Mastrangelo MJ. Immunopharmacologic analysis of an autologous, hapten-modified human melanoma vaccine. *J Clin Oncol*. 2004 Feb 1;22(3):403-15.
- [6] CIMAvax EGF® Conjugado químico de Factor de Crecimiento Epidérmico humano recombinante acoplado a la proteína recombinante rP64K, Número de Registro B- 08- 063- L03- C, Centro de Inmunología Molecular (CIM), Cuba. Centro para el Control de la Calidad de los Medicamentos (CECMED). Havana, Cuba; 2008 June 12.
- [7] Couch M, Saunders JK, O'Malley BW Jr, Pardoll D, Jaffee E. Spatial distribution of tumor vaccine improves efficacy. *Laryngoscope*. 2003 Aug;113(8):1401-5.
- [8] Crombet T, Neninger E, Catalá M, García B, Leonard I, Martínez L, et al. Treatment of NSCLC Patients with an EGF-Based Cancer Vaccine: report of a Phase I trial. *Cancer Biol Ther*. 2006 Feb;5(2):130-40.
- [9] Crombet T, Rodríguez PC, Neninger E, García B, Lage A. CIMAvax EGF (EGF-P64K) vaccine for the treatment of non-small-cell lung cancer. *Expert Rev Vaccines*. 2015 Oct;14(10):1303-11.
- [10] Cupps TR, Edgar LC, Fauci AS. Suppression of human B lymphocyte function by cyclophosphamide. *J Immunol*. 1982 Jun;128(6):2453-7.
- [11] Dass S, Vital EM, Emery P. Development of psoriasis after B cell depletion with rituximab. *Arthritis Rheum*. 2007 Aug;56(8):2715-8.
- [12] del Cristo IS, Fors MM, Viada C, Guerra PP, Hernández A, Hernández M. Experiencias de un ensayo clínico en la atención primaria de salud. *Revista Bionatura*. 2015; 1(1):20-30.
- [13] Dunn GP, Old LJ, Schreiber RD. The Three Es of Cancer Immunoediting. *Annu Rev Immunol*. 2004;22:329-60.
- [14] Ehrke MJ. Immunomodulation in cancer therapeutics. *Int Immunopharmacol*. 2003 Aug;3(8):1105-19. Review.
- [15] Emens LA, Jaffee EM. Leveraging the activity of tumor vaccines with cytotoxic chemotherapy. *Cancer Res*. 2005 Sep 15;65(18):8059-64.
- [16] Fine GD. Consequences of delayed treatment effects on analysis of time-to-event endpoints. *Drug Inf J*. 2007;41(4):535-9.
- [17] Galán Y, Fernández L, Torres P, García M. Trends in Cuba's Cancer Incidence (1990 to 2003) and Mortality (1990 to 2007). *MEDICC Review*. 2009;11(3):19-26.
- [18] García B, Neninger E, de la Torre A, Leonard I, Martínez R, Viada C, et al. Effective Inhibition of the Epidermal Growth Factor/Epidermal Growth Factor Receptor Binding by Anti-Epidermal Growth Factor Antibodies Is Related to Better Survival in Advanced Non Small-Cell Lung Cancer Patients Treated with the

- Epidermal Growth Factor Cancer Vaccine. *Clin Cancer Res.* 2008 Feb 1;14(3):840–6.
- [19] Gardiner DF, Huang Y, Basu S, Leung L, Song Y, Chen Z, et al. Multiple-site DNA vaccination enhances immune responses in mice. *Vaccine.* 2006 Jan 16;24(3):287–92.
- [20] Gerard C, Debruyne C. Immunotherapy in the landscape of new targeted treatments for non-small cell lung cancer. *Mol Oncol.* 2009 Dec;3(5–6):409–24.
- [21] Ghiringhelli F, Larmonier N, Schmitt E, Parcellier A, Cathelin D, Garrido C, et al. CD4+CD25+ regulatory T cells suppress tumor immunity but are sensitive to cyclophosphamide which allows immunotherapy of established tumors to be curative. *Eur J Immunol.* 2004 Feb;34(2):336–44.
- [22] Goetz M, Atreya R, Ghalibaan M, Galle PR, Neurath MF. Exacerbation of ulcerative colitis after rituximab salvage therapy. In—— am Bowel Dis. 2007 Nov;13(11):1365–8.
- [23] González G, Crombet T, Catalá M, Mirabal V, Hernández JC, González Y, et al. A novel cancer vaccine composed of human-recombinant epidermal growth factor linked to a carrier protein: report of a pilot clinical trial. *Ann Oncol.* 1998 Apr;9(4):431–5.
- [24] González G, Crombet T, Neninger E, Viada C, Lage A. Therapeutic vaccination with epidermal growth factor (EGF) in advanced lung cancer: analysis of pooled data from three clinical trials. *Hum Vaccin.* 2007 Jan-Feb;3(1):8–13.
- [25] González G, Crombet T, Torres F, Catalá M, Alfonso L, Osorio M, et al. Epidermal growth factor-based cancer vaccine for non-small-cell lung cancer therapy. *Ann Oncol.* 2003 Mar;14(3):461–6.
- [26] González G, Lage A. Cancer Vaccines for Hormone/Growth Factor Immune Deprivation: A Feasible Approach for Cancer Treatment. *Curr Cancer Drug Targets.* 2007 May;7(3):229–41.
- [27] González G, Pardo OL, Sánchez B, García JL, Beausoleil I, Marinello P, et al. Induction of immune recognition of self-epidermal growth factor II: characterization of the antibody immune response and the use of a fusion protein. *Vac Res.* 1997;6(2):91–100.
- [28] González G, Sánchez B, Suárez E, Beausoleil I, Perez O, Lastre M, et al. Induction of immune recognition of self-epidermal growth factor (EGF): effect on EGF biodistribution and tumor growth. *Vac Res.* 1996;5(4):233–44.
- [29] Hanahan D, Weinberg RA. The Hallmarks of cancer. *Cell.* 2000 Jan 7;100(1):57–70. Review.
- [30] Hernández AM, Vázquez AM. Racotumomab-alum vaccine for the treatment of non-small-cell lung cancer. *Expert Rev Vaccines.* 2015 Jan;14(1):9–20.
- [31] Hirsch FR, Herbst RS, Olsen C, Chansky K, Crowley J, Kelly K, et al. Increased EGFR Gene Copy Number Detected by Fluorescent In Situ Hybridization Predicts Outcome in Non-Small-Cell Lung Cancer Patients Treated With Cetuximab and Chemotherapy. *J Clin Oncol.* 2008 Jul 10;26(20):3351–7.
- [32] Hirsch FR, Varella-García M, Cappuzzo F. Predictive value of EGFR and HER2 overexpression in advanced non-small-cell lung cancer. *Oncogene.* 2009 Aug;28 Suppl 1:S32–7.
- [33] Hoos A, Parmiani G, Hege K, Sznol M, Loibner H, Eggermont A, et al. A Clinical Development Paradigm for Cancer Vaccines and Related Biologics. *J Immunother.* 2007 Jan;30(1):1–15.
- [34] Kroemer G, Pouyssegur J. Tumor Cell Metabolism: Cancer's Achilles' Heel. *Cancer Cell.* 2008 Jun;13(6):472–82. Review.
- [35] Lage A, Crombet T, González G. Targeting epidermal growth factor receptor signaling: early results and future trends in oncology. *Ann Med.* 2003;35(5):327–36.
- [36] Lage A, Pérez R, Fernández LE. Therapeutic cancer vaccines: at midway between immunology and pharmacology. *Curr Cancer Drug Targets.* 2005 Dec;5(8):611–27.
- [37] Lage A. Connecting immunology research to public health: Cuban biotechnology. *Nature Immunology.* 2008;9(2):109–12.
- [38] Lage A. Transforming Cancer Indicators Begs Bold New Strategies from Biotechnology. *MEDICC Review.* 2009;11(3):8–12.
- [39] Macías A, Azavedo E, Hägerström T, Klintonberg C, Pérez R, Skoog L. Prognostic significance of the receptor for epidermal growth factor in human mammary carcinoma. *Anticancer Res.* 1987 May–Jun;7(3 Pt B):459–64.
- [40] Madan RA, Mohebtash M, Schlom J, Gulley JL. Therapeutic vaccines in metastatic castration-resistant prostate cancer: principles in clinical trial design. *Expert Opin Biol Ther.* 2009 Oct 27;10(1):1–10.
- [41] Man S, Bocci G, Francia G, Green SK, Jothy S, Hanahan D, et al. Antitumor effects in mice of low-dose (metronomic) cyclophosphamide administered continuously through the drinking water. *Cancer Res.* 2002 May 15;62(10):2731–5.
- [42] Mantovani A, Allavena P, Sica A, Balkwill F. Cancer-related in—— amma-tion. *Nature.* 2008 Jul 24;454(7203):436–44. Review.
- [43] Mantovani A. Inflaming Metastasis. *Nature.* 2009 Jan 1;457(7225):36–7.

- [44] Maurenis Hernandez, Ramon A Ortiz, Elia Neninger, Rosa M Amador, Mireisy Cala, et. Al. P2.40: CIMAvaxEGF Vaccine for the Treatment of Real-World NSCLC Patients: Track: Immunotherapy. *J Thorac Oncol* 2016 Oct 22;11(10S):S242. Epub 2016 Dec 22.
- [45] Mazorra Z, Crombet T. Pilot study of a novel combination of two therapeutic vaccines in advanced non-small-cell lung cancer patients *Cancer Immunology, Immunotherapy* July 2014, Volume 63, Issue 7, pp 737–747
- [46] Montero E, Valdés M, Avellanet J, López A, Pérez R, Lage A. Chemotherapy induced transient B-cell depletion boosts antibody-forming cells expansion driven by an epidermal growth factor-based cancer vaccine. *Vaccine*. 2009 Apr 6;27(16):2230–9.
- [47] Neninger E, Verdecia BG, Crombet T, Viada C, Pereda S, Leonard I, et al. Combining an EGF-based cancer vaccine with chemotherapy in advanced nonsmall cell lung cancer. *J Immunother*. 2009 Jan;32(1):92–9.
- [48] Neninger Vinageras E, de la Torre A, Osorio Rodríguez M, Catalá Ferrer M, Bravo I, Mendoza del Pino M, et al. Phase II randomized controlled trial of an epidermal growth factor vaccine in advanced NSCLC. *J Clin Oncol*. 2008 Mar 20;26(9):1452–8.
- [49] Pérez L, Estévez D, Gastón Y, Macías A, Viada CE. Seguridad del Racotumomab en el tratamiento de pacientes con cáncer de pulmón de células no pequeñas. *VacciMonitor* 2013; 22(1):10-14.
- [50] Pérez R, Pascual M, Macías A, Lage A. Epidermal growth factor receptors in human breast cancer. *Breast Cancer Res Treat*. 1984 Sep;4(3):189–93.
- [51] Rodríguez G, González G, Crombet T, Lage A. CIMAvax EGF Development and Manufacturing. *Biopharm International*. 2008 Oct; Suppl 1:S36–42.
- [52] Rodríguez PC, González I, González A, Avellanet J, López A, Pérez R, et al. Priming and boosting determinants on the antibody response to an Epidermal Growth Factor-based cancer vaccine. *Vaccine*. 2008 Aug 26;26(36):4647–54.
- [53] Rodriguez PC, Popa X, Martinez O, Mendoza S, Santiesteban E, et. al. A Phase III Clinical Trial of the Epidermal Growth Factor Vaccine CIMAvax-EGF as Switch Maintenance Therapy in Advanced Non-Small-Cell Lung Cancer Patients. *Clin Cancer Res* 2016 Feb 29. Epub 2016 Feb 29.
- [54] Romero T. Changing the Paradigm of Cancer Control in Cuba. *MEDICC Review*. 2009;11(3):5–7.
- [55] Santiesteban E, Alfonso S, Neninger E, Acosta S, Pérez L, Macías A, et al. Racotumomab anti-idiotypic vaccine for patients with non-small cell lung cancer. *International Journal of Clinical Medicine* 2014; 5: 844-850.
- [56] Schlom J, Arlen PM, Gulley JL. Cancer Vaccines: Moving Beyond Current Paradigms. *Clin Cancer Res*. 2007 Jul 1;13(13):3776–82.
- [57] Sinkovics JG, Horvath JC. Evidence accumulating in support of cancer vaccines combined with chemotherapy: a pragmatic review of past and present efforts. *Int J Oncol*. 2006 Oct;29(4):765–77.
- [58] Starzl TE, Zinkernagel RM. Antigen localization and migration in immunity and tolerance. *N Engl J Med*. 1998 Dec 24;339(26):1905–13.
- [59] Terando A, Mulé JJ. On combining antineoplastic drugs with tumor vaccines. *Cancer Immunol Immunother*. 2003 Nov;52(11):680–5.
- [60] Toffoli G, De Mattia E, Cecchin E, Biason P, Masier S, Corona G. Pharmacology of epidermal growth factor inhibitors. *Int J Biol Markers*. 2007 Jan–Mar;22(1 Suppl 4):S24–39.
- [61] Viada C, Dupuy JF, Fors M, Luaces P, González G, Neninger E, et al. CIMAvax®EGF vaccine therapy for non-small cell lung cancer: A weighted log-rank tests-based evaluation. *Modern Chemotherapy* 2013, 2(3):51-56.
- [62] Viada C, Dupuy JF, Fors M, Rodríguez PC, González G, et al. Stratified Cox Regression Analysis of Survival under CIMAvax®EGF Vaccine. *Journal of Cancer Therapy*, 2013, 4, 8-14.
- [63] Viada C, Fors M, Neninger E, Alfonso S, Santiesteban E, Mendoza I, y col. Seguridad de la vacuna anti-idiotípica 1E10 en pacientes con tumores de diversas localizaciones. *Revista Bionatura*. 2015; 1(1):31-43.
- [64] Zinkernagel RM, Hengartner H. Regulation of the immune response by antigen. *Science*. 2001 Jul 13;293(5528):251–3.

APLICACIÓN REAL DE LA TÉCNICA DE HORVITZ DE RESPUESTA ALEATORIZADA AL ACOSO SEXUAL

B. Cobo¹, F. López –Torrecillas², D. Molina³

¹Departamento de Estadística e Investigación Operativa, Facultad de Ciencias. Universidad de Granada. España.

²Departamento de Personalidad, Evaluación y Tratamiento Psicológico, Facultad de Psicología. Universidad de Granada. España.

³Departamento de Estadística e Investigación Operativa, Facultad de Educación, Economía y Tecnología. Universidad de Granada. Campus de Ceuta.

ABSTRACT

In this paper we study the sexual harassment among students of the University of Granada. Sexual harassment is a sensitive issue for the population, so if we make direct surveys, we will find problems, such as social desirability, which is the tendency of respondents to answer in terms of what that is socially acceptable, that respondents do not answer truthfully because they feel their privacy violated or simply they refuse to answer. To protect the anonymity of the respondent and reduce the risk of evasion or lack of response to sensitive questions are used randomized response techniques.

The aim of the randomized response techniques and indirect questioning is to maintain the privacy of respondents. Using these new techniques we obtain estimations more accurate than those obtained by direct response.

KEY WORDS: Sexual harassment, Randomized response, Sampling, Confidentiality

RESUMEN

En este trabajo estudiamos el acoso sexual en alumnos de la Universidad de Granada. El acoso sexual es un tema sensible para la población, por lo que si realizamos las encuestas de forma directa, nos vamos a encontrar con problemas, como por ejemplo, la deseabilidad social, que es la tendencia de los encuestados a responder en función de lo que es socialmente aceptable, que los entrevistados no contesten verazmente porque vean violada su intimidad o simplemente que se nieguen a responder. Para proteger el anonimato del encuestado y reducir el riesgo de evasión o falta de respuesta a las preguntas sensibles utilizamos las técnicas de respuesta aleatoria.

Las técnicas de respuesta aleatoria y las preguntas indirectas tienen como objetivo mantener la privacidad de los encuestados. Mediante estas nuevas técnicas se obtienen estimadores que son más precisos en comparación a respuesta directa.

1. INTRODUCCIÓN

En nuestro estudio vamos a realizar una encuesta para investigar el acoso sexual en la Universidad de Granada.

El acoso sexual es un tipo de violencia, que se produce en el entorno laboral, un comportamiento verbal, psicológico no deseado, con el propósito de atentar contra la dignidad o crear un entorno intimidatorio, hostil, humillante u ofensivo.

El acoso sexual es una forma de discriminación por razón de género, si bien los hombres pueden ser también objeto de acoso sexual, la realidad es que la mayoría de víctimas son mujeres. Alguno de los efectos que sufren las víctimas de este tipo de acoso son: estrés emocional, humillación, ansiedad, depresión, ira, impotencia, fatiga, tensión en el trabajo, disminución en productividad, bajo rendimiento, absentismo, etc...

Estudios realizados en España (Muñoz-Rivas, Graña, O'Leary & González 2007; 2009) muestran un porcentaje significativamente mayor de mujeres que participan en agresiones verbales (95,3% vs. 92,8%), mientras que los varones se involucran en agresiones físicas más severas (4.6% vs. 2.0%) y producen peores consecuencias para la salud de sus compañeras

El acoso sexual es un tema muy sensible en la sociedad, por lo que los encuestados suelen mentir o no contestar ante tales cuestiones debido a que ven violada su privacidad al tener que responder ante un entrevistador. En este tipo de encuesta además surge un problema denominado deseabilidad social, que se define como la tendencia por parte de los encuestados a responder en función de lo que es socialmente aceptable. Esto provoca que las estimaciones obtenidas se vean alteradas y pierdan confiabilidad.

Para evitar estos problemas, Warner en 1965 propuso las técnicas de respuesta aleatoria con el objetivo de proteger la privacidad de los encuestados en el estudio de temas sensibles y de este modo evitar el sesgo de respuesta.

El método original de Warner consiste en la utilización de un mecanismo aleatorio por medio del cual se selecciona una de dos preguntas complementarias: ¿pertenece al grupo con la característica A ? o ¿pertenece al grupo que no tiene la característica A ?, donde A es la característica sensible de interés. El entrevistado contestará sí o no y el entrevistador no tiene la posibilidad de saber qué pregunta contestó el entrevistado.

En los últimos años son muchos los autores que utilizan estas técnicas para obtener respuestas veraces ante preguntas sensibles. Por ejemplo, consumo de drogas (Goodstadt & Gruson, 1975), el impacto de la infección por el VIH/SIDA en Bostwana (Arnab & Singh, 2010), uso de drogas ilícitas y recetadas para el mejoramiento cognitivo y del estado de ánimo entre los cirujanos (Franke, Bagusat, Dietz, Hoffmann, Simon, Ulrich & Lieb, 2013), y el aborto inducido y el estatus de inmigrantes irregulares entre mujeres extranjeras en Calabria (Perri, Pelle & Stranges, 2015).

Para el tema específico del comportamiento sexual, LaBrie & Earleywine (2000) y Walsh & Braithwaite (2008) usan métodos de interrogación indirecta para investigar la actividad sexual de riesgo.

De Jong, Pieters & Stremersch, (2012) comparan muestras de 2 países sobre actitudes sexuales permisivas y comportamientos sexuales de riesgo usando un RRT.

En este capítulo se verá el desarrollo y los resultados de una encuesta utilizando la el modelo de Horvitz cuando la proporción de personas que tienen la característica inocua es conocida, para estudiar el acoso sexual.

2. PARTICIPANTES Y MÉTODOS DE MUESTREO

La población objetivo de esta encuesta fueron los estudiantes de la Universidad de Granada (España). Los estudiantes fueron seleccionadas utilizando un diseño de muestreo complejo, en el que se extrajo una muestra por conglomerados estratificados (por facultad) de forma que todas las asignaturas y cursos fueron representados en proporción al total del número de estudiantes. Todos los cuestionarios se pasaron durante horario de clase. Todos los estudiantes fueron invitados a participar en el estudio y proporcionaron su consentimiento mediante firma.

3. PROCEDIMIENTO Y MEDIDA

El cuestionario comenzó con un conjunto de preguntas demográficas básicas, seguidas de algunas preguntas académicas, y luego una serie de ítems referidos a comportamientos sensibles, tales como copia en exámenes, consumo de drogas y acoso sexual.

Para aleatorizar las respuestas, se utilizó el modelo propuesto por Horvitz, Shah & Simmons, (1967) y extendido por Greenberg, Abul-Ela, Simmons & Horvitz (1969). En comparación con variantes alternativas de las técnicas de respuesta aleatoria, este diseño es menos complejo y muy adecuado.

El mecanismo de aleatorización utilizado fue la aplicación “Randomizers” (Play Store, 2015) con la opción “Coin Flipper”, que se instaló previamente en el teléfono de los estudiantes. La aplicación es muy sencilla de utilizar: el usuario toca el botón “Randomize” y se muestra un lado de la moneda – cara o cruz.

En nuestro estudio, si aparece “cara” el estudiante responde la pregunta principal (“¿Alguna vez has acosado sexualmente a alguien?”), y si aparece “cruz”, responde la pregunta alternativa (“¿Tu DNI termina en 5?”)

La figura 1 muestra la interfaz de la aplicación y la figura 2 muestra el procedimiento de respuesta seguido por el encuestado.



Figura 1: aplicación “Randomizers”

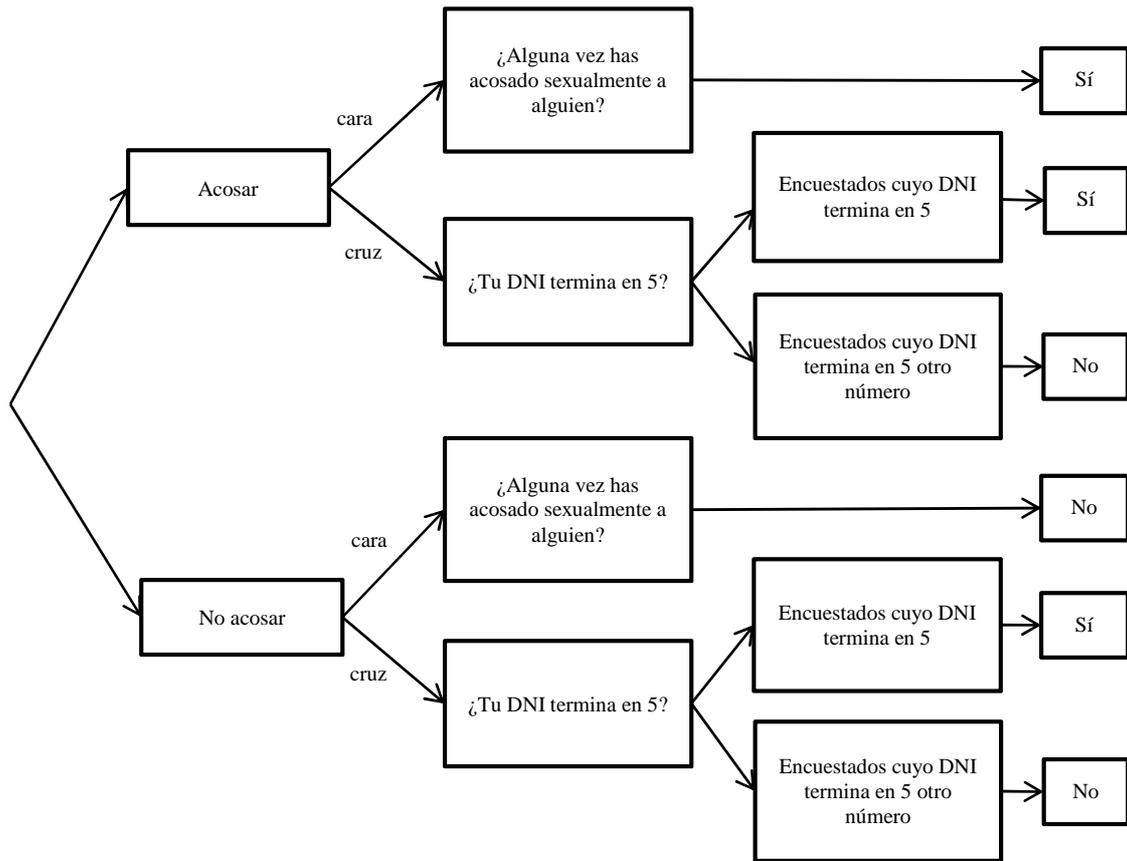


Figura 2: procedimiento de respuesta del encuestado

El profesor explicó que esta técnica preservaba el anonimato de los estudiantes con el objetivo de no provocar desconfianza en ellos y todos los estudiantes completaron el cuestionario completo.

La recogida de datos y el trabajo de campo fueron realizados por el grupo de investigación FQM365 del Plan Andaluz de Investigación. Ocho entrevistadores llevaron a cabo las entrevistas de enero a marzo de 2015 en España. Los datos se obtuvieron de 710 estudiantes.

4. ANÁLISIS ESTADÍSTICO

Los pesos del diseño se calcularon mediante un diseño aleatorio estratificado y se modificaron para ajustar el sesgo de cobertura. Todos los análisis estadísticos se realizaron utilizando los pesos de muestreo.

La diferencia entre las medias se probó usando las pruebas t, y la diferencia entre las proporciones se probó mediante la prueba de chi-cuadrado. El nivel de significación se estableció en 0,05.

El estimador insesgado propuesto por Horvitz, Shah, et al. (1967) y ampliado por Greenberg, Abul-Ela, et al. (1969) se utilizó para estimar los valores medios de la variable de interés.

El modelo de Horvitz es el siguiente: Sea z_i la respuesta aleatoria obtenida del i -ésimo alumno. Definimos $z_i = 1$, si la respuesta del estudiante i es "Sí" y $z_i = 0$ si la respuesta es "No". Sea Q la proporción de estudiantes cuyo DNI termina en 5. Definimos la variable transformada como

$$r_i = \frac{z_i - (1 - p)Q}{p}$$

siendo $p = 0.5$ y $Q = 1/10$; la varianza de r_i es

$$\hat{\phi}_i = r_i(r_i - 1)$$

Un estimador insesgado de la proporción para la encuesta RR es

$$\hat{Y}_{ht}(r) = \frac{1}{N} \sum_{i \in S} \frac{r_i}{\pi_i} \quad (1)$$

y el estimador insesgado de la varianza de este estimador es

$$\hat{V}_{ht}(r) = \frac{1}{N^2} \left[\frac{1}{2} \sum_{i \neq j} \sum_{j \in S} \frac{(\pi_i \pi_j - \pi_{ij})}{\pi_{ij}} \left(\frac{r_i}{\pi_i} - \frac{r_j}{\pi_j} \right)^2 + \sum_{i \in S} \frac{\hat{\phi}_i}{\pi_i} \right] \quad (2)$$

donde π_i y π_{ij} son las probabilidades de inclusión (Singh, 2003) de la i -ésima unidad y la i -ésima y j -ésima unidades ($i \neq j$) respectivamente. Las probabilidades de inclusión de primer y segundo orden se calculan a partir de muestreo aleatorio por conglomerados estratificados.

El intervalo de confianza al $(1 - \alpha)\%$ es dado por

$$ci = \left(\hat{Y}_{ht}(r) - z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{ht}(r)}, \hat{Y}_{ht}(r) + z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}_{ht}(r)} \right)$$

donde $z_{1-\frac{\alpha}{2}}$ denota el $(1 - \alpha)\%$ cuantil de una distribución normal estándar.

Todos los análisis estadísticos se realizaron utilizando el software R. Utilizamos un paquete específico para manejar los datos de respuesta aleatoria obtenidos de encuestas complejas (RRTCS, Rueda, Cobo & Arcos, 2015). Concretamente, en este paquete utilizamos la función Horvitz ().

5. RESULTADOS

La distribución sociodemográfica de la muestra se muestra en la tabla 1.

	<i>Frecuencia</i>	<i>Porcentaje</i>
Total	710	100%
Sexo		
Masculino	335	47.15%
Femenino	375	52.85%
Facultad		
Ciencias e Ingeniería	225	31.7%
Ciencias de la Salud	154	21.7%
Ciencias Sociales y Derecho	331	46.6%

Tabla 1. Distribución sociodemográfica de la muestra

La encuesta se compuso de 710 estudiantes, de los cuales el 52,85% eran mujeres y el 47,15% hombres. Por Facultades, el 31,7% de los estudiantes estaba en la Facultad de Ciencias e Ingeniería, el 21,7% en la Facultad de Ciencias de la Salud y el 46,6% en la Facultad de Ciencias Sociales y Derecho.

La prevalencia de los estudiantes que respondieron positivamente a la pregunta sensible se resume en la Tabla 2.

	Estimación	Desviación estándar	Intervalo de confianza (95%)	
			Límite inferior	Límite superior
Total	0.1281	0.0234	0.0817	0.1745
Sexo				
Masculino	0.1794	0.0370	0.1166	0.2521
Femenino	0.0589	0.0273	0.0052	0.1126
Facultad				
Ciencias e Ingeniería	0.1400	0.0427	0.0560	0.2239
Ciencias de la Salud	0.1077	0.0486	0.0122	0.2033
Ciencias Sociales y Derecho	0.1296	0.0346	0.0615	0.1977

Tabla 2. Prevalencia del acoso sexual

La prevalencia estimada de estudiantes que habían acosado sexualmente a alguien varió de 8,17% a 17,45% (al nivel de confianza del 95%).

La proporción de mujeres que habían acosado sexualmente es del 6,89%, mucho menor que la proporción de hombres, siendo el 17,94%. Estas diferencias son significativas ($p_value = 0.003959$)

Si nos fijamos en la prevalencia estimada por facultad, vemos que los valores son muy similares. Las diferencias por facultades no son significativas ($p_value = 0.888$)

REFERENCIAS

- [1] ARNAB, R. & SINGH, S. (2010). Randomized response techniques: An application to the Botswana AIDS impact survey. **Journal of Statistical Planning and Inference**, 140, 941- 953. <http://dx.doi.org/10.1016/j.jspi.2009.09.019>.
- [2] DE JONG, M.G., PIETERS, R. & STREMERSCHE, S. (2012). Analysis of sensitive questions across cultures: An application of multigroup item randomized response theory to sexual attitudes and behavior. **Journal of Personality and Social Psychology**, 103,543-564. [http://dx. doi: 10.1037/a0029394](http://dx.doi.org/10.1037/a0029394).
- [3] FRANKE, A.G., BAGUSAT, C., DIETZ P., HOFFMANN I., SIMON P., ULRICH R., LIEB K. (2013). Use of illicit and prescription drugs for cognitive or mood enhancement among surgeons. **BMC Medicine**. DOI: 10.1186/1741-7015-11-102
- [4] GREENBERG, B.G., ABUL-ELA, A.L., SIMMONS, W.R., HORVITZ, D.G. (1969). The unrelated question RR model: Theoretical framework. **Journal of the American Statistical Association**, 64, 520-539.
- [5] GOODSTADT, M.S. & GRUSON, V. (1975). The randomized response technique: A test on drug use. **Journal of the American Statistical Association**, 70, 814-818. <http://dxdoi:10.1017/S037689290999004X>
- [6] HORVITZ, D.G., SHAH, B.V., SIMMONS, W.R. (1967). The unrelated question RR model. **Proceedings of the Social Statistics Section of the American Statistical Association**. 65-72. Alexandria, VA: ASA.
- [7] LABRIE, J. W. & EARLEYWINE, M. (2000). Sexual Risk Behaviors and Alcohol: Higher Base Rates Revealed Using the Unmatched-Count Technique. **Journal of Sex Research**, 37(4), 321–326. doi:[http://dx.doi:10.1080/00224490009552054](http://dx.doi.org/10.1080/00224490009552054).
- [8] MUÑOZ-RIVAS, M.J., GRAÑA, J.L., O'LEARY, K.D. & GONZÁLEZ, P. (2007). Aggression in adolescent dating relationships: Prevalence, Justification and Health Consequences. **Journal of Adolescent Health**, 40, 298-304. doi:[http://dx.doi:10.1016/j.jadohealth.2006.11.137](http://dx.doi.org/10.1016/j.jadohealth.2006.11.137).
- [9] MUÑOZ-RIVAS, M.J., GRAÑA, J.L., O'LEARY, K.D. & GONZÁLEZ, P. (2009). Prevalence and predictors of sexual aggression in dating relationships of adolescents and young adults. **Psicothema**, 21(2), 234-240. Retrieved from <http://www.psicothema.com/pdf/3620.pdf>.
- [10] PLAY STORE (2015). Download the application “Baraja Española”. <https://play.google.com/store/apps/details?id=cartas.barajaes>
- [11] PERRI, P.F., PELLE, E. & STRANGES, M. (2015). Estimating induced abortion and foreign irregular presence using the randomized response crossed model. **Social Indicators Research**, 1-18. Published online. [http://dx.doi: 10.1007/s11205-015-1136-x](http://dx.doi.org/10.1007/s11205-015-1136-x).
- [12] RUEDA, M., COBO, B., ARCOS, A. (2015). Package ‘RRTCS’: Randomized Response Techniques for Complex Surveys. URL = <http://cran.r-project.org/web/packages/RRTCS/>

- [13] WALSH, J. A. & BRAITHWAITE, J. (2008). Self-Reported Alcohol Consumption and Sexual Behavior in Males and Females: Using the Unmatched-Count Technique to Examine Reporting Practices of Socially Sensitive Subjects in a Sample of University Students. **Journal of Alcohol and Drug Education**, 52(2), 49–72. <https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=245966>
- [14] WARNER, S.L. (1965). Randomized Response: a survey technique for eliminating evasive answer bias. **Journal of the American Statistical Association** 60, 63-69.