Desarrollo de nuevos modelos y métodos matemáticos para la toma de decisiones:

La pandemia y su impacto en la sociedad

Editores

Carlos N. Bouza Herrera

Agustín Santiago Moreno

José Maclovio Sautto Vallejo

Red Iberoamericana de Estudios Cuantitativos Aplicados

Desarrollo de nuevos modelos y métodos matemáticos para la toma de decisiones:

La pandemia y su impacto en la sociedad

Serie Métodos cuantitativos aplicados

Editores:
Carlos N. Bouza Herrera
Agustín Santiago Moreno
José M. Sautto Vallejo

Ignacio Mariscal 59-22
Col. Tabacalera Cuauhtémoc
C.P. 06030

Esta publicación se puede descargar de forma gratuita.

ISBN: 978-607-59449-2-0

Primera edición Marzo de 2023

Índice de capítulos

Nombre	**Afiliación**
Agustín Santiago Moreno	Facultad de Matemáticas, Universidad Autónoma de Guerrero, México.
Aidé Ibárez Castro	Secretaría de Salud, Guerrero, México.
América A. Rosales Ibárez	Secretaría de Salud, Guerrero, México.
Antonio Rosales Jiménez	Universidad Autónoma de Guerrero, Facultad de Medicina.
Carlos Alberto Álvarez Bravo	Dpto. de Matemática, Facultad de Ingeniería, Universidad de Cienfuegos, Cienfuegos, Cuba
Carlos N. Bouza Herrera	Facultad de Matemática y computación, Universidad de la Habana
Cristina Mendoza Rodríguez	Universidad de ciencias médicas de Villa Clara, Cuba, Facultad de medicina.
Faizan Danish	Research Consultation Services, Doha, Qatar
Gladys Linares Fleites	Posgrado de Ciencias Ambientales. Instituto de Ciencias. Benemérita Universidad Autónoma de Puebla. Edificio IC3 118, Ciudad Universitaria, Puebla, México
Ignacio Caamal Cauich	Universidad Autónoma de Chapingo (UACH). Escuela de Economía
Jenny Paola Lis Gutiérrez	Fundación Universitaria Konrad Lorenz (Bogotá, Colombia)
José A. Betancourt Bethencourt	Universidad de Ciencias Médicas Carlos J. Finlay, Cuba
José Félix García Rodríguez	Universidad Juárez Autónoma de Tabasco (UJAT). División Académica de Ciencias Económico Administrativas (DACEA)
José M. Sautto Vallejo	Facultad de Matemáticas, Universidad Autónoma de Guerrero
Juan Felipe Medina Mendieta	Universidad de Cienfuegos “Carlos Rafael Rodríguez”, Cuba.
Lourdes del Carmen Pineda Celaya	Universidad Popular Autónoma del Estado de Puebla
Luis Manuel Hernández Govea	Universidad Juárez Autónoma de Tabasco. UJAT
Luis René Marcial Castillo	Benemérita Universidad Autónoma de Puebla, Puebla, México
Manuel Cortés Cortés	Universidad de Cienfuegos “Carlos Rafael Rodríguez”, Cuba.
Marcela Rivera Martínez	Benemérita Universidad Autónoma de Puebla, Puebla, México
María de Lourdes Sandoval Solís	Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla. Edificio CCO2 106, Ciudad Universitaria, Puebla, México
Mario Moreira	Ingeniero de Proceso y Mejora Continua, Cementos Cienfuegos S.A, Cienfuegos, Cuba
Naamán Izquierdo Balcázar	Universidad Juárez Autónoma de Tabasco. UJAT
Octaviano Juárez Romero	Universidad Autónoma de Guerrero, México
Oscar Rene Salgado Guzmán	Benemérita Universidad Autónoma de Puebla, Puebla, México
Pablo Otoniel Juárez Moreno	Universidad Autónoma de Guerrero, México
S.E.H. Rizvi	Division of Statistics and Computer Science, Faculty of Basic Sciences, SKUAST-JAMMU, (J&K), India
Sira M. Allende Alonso	Facultad de Matemática y computación, Universidad de la Habana
Verna Gricel Pat Fernández	Universidad Autónoma Chapingo, México.

Prólogo

Durante los momentos más aciagos de la pandemia de coronavirus que padeció México y el mundo, los distintos grupos de investigación académica continuaron con su meritoria actividad, aún con las limitaciones que el entorno les imponía, o aun utilizando la pandemia como una oportunidad para contribuir a la solución o comprensión del problema. La red Iberoamericana de Estudios Cuantitativos Aplicados (RIDECA), entrega a la comunidad académica del mundo el tomo 2 del libro "Desarrollo de nuevos modelos y métodos matemáticos para la toma de decisiones", en él se puede apreciar la creatividad y la aplicación de distintos modelos a distintos casos. Regresión logística y curva de Gompertz para el estudio de covid-19, Nuevo escenario de vulnerabilidad y pobreza a causa de covid-19, bajo el enfoque de Planeación estratégica para el impulso del desarrollo local en México, la selección de muestra de poblaciones con estructura de red: modelos alternativos para evaluar aspectos de una pandemia, Vinculación del programa r con googlemap para análisis espaciales de covid19, Algunos elementos sobre las curvas roc: teoría y herramientas, que puedes ser utilizadas en el estudios de las secuelas de los recuperados por Covid19, Método para determinar las fases minerales del clínker y su influencia en reducir los daños al medio ambiente, Caracterización de los indicadores agrarios de la producción de limón persa y Diagnóstico de depresión en adultos mayores en el nivel primario de atención mediante el uso de un modelo de regresión logística binaria..

Al leer este libro, uno puede apreciar la potencialidad de la modelación matemática para la comprensión de fenómenos, tales como la pandemia del coronavirus y otros tantos que puedes ser susceptibles de modelación, dadas las regularidades de su comportamiento.

Me place felicitar a la Red Iberoamericana de Estudios Cuantitativos Aplicados (RIDECA), por su esfuerzo y contribución para la comprensión de este tipo de fenómenos, como son las pandemias. Enhorabuena.

Dra. Aidé Ibárez Castro

Secretaria de Salud del Estado de Guerrero, México

Marzo de 2023

Capítulo 1 pp 1-12

REGRESIÓN LOGÍSTICA Y CURVA DE GOMPERTZ PARA LA COVID-19. CASO CUBA.

Juan Felipe Medina Mendieta[1] y Manuel Cortés Cortés[1]

[1]Universidad de Cienfuegos "Carlos Rafael Rodríguez", Cuba.

RESUMEN

El planeta se encuentra en la actualidad en una situación desfavorable de salud. La Organización Mundial de la Salud declara el 11 de marzo del 2020 pandemia de la Covid-19, instando a la imposición de medidas por parte de los gobiernos de cada país. Para estos gobiernos es de vital importancia contar con estimados de casos contagiados y fallecidos para poder aplicar las medidas necesarias con los recursos que tienen a su disposición.

El objetivo fundamental del presente trabajo es obtener predicciones para los picos de casos confirmados y fallecidos en Cuba, así como estimaciones de la cantidad total de los mismos. Para ello, mediante el ajuste de curvas del tipo logística y curva de Gompertz, aplicando el método estadístico de los mínimos cuadrados y haciendo uso de herramientas informáticas se estudia el crecimiento de casos contagiados y de fallecidos. Se estudian los casos de Italia y España debido a que han pasado el pico de la epidemia y a que Cuba ha utilizado una variante de confinamiento similar a la utilizada por estos países.

Como resultados fundamentales se tiene que existe una adecuación de los modelos con respecto a los valores pronosticados y los reales, correspondientes a los casos de Italia y España, lo cual permite una confiabilidad de los mismos para los pronósticos efectuados para Cuba.

Las predicciones realizadas para Cuba plantean que los picos estimados de contagios y fallecidos se alcanza entre el 14 - 19 de abril y el 18 de abril - 5 de mayo respectivamente. Se pronostica un total de contagiados entre 1 482 a 2 678 y fallecidos entre 63 a 211.

Palabras claves: Covid-19, modelos predictivos, ajuste de curvas, Cuba.

LOGISTIC REGRESSION AND GOMPERTZ CURVE FOR COVID-19. CUBA CASE.

ABSTRACT

The planet is currently in a difficult health situation. The World Health Organization declares a Covid-19 pandemic on March 11, 2020. This organization has indicated to take measures to the governments of each country. For this, it is important to know estimates for infected and deceased cases.

The main objective of this work is to obtain predictions for the peaks of confirmed and deceased cases in Cuba and their totals. For this, the method of least squares was used for the adjustment of logistics-type curves and the Gompertz curve and the use of computer tools. The growth of infected and deceased cases respectively in

[1] jfelipemm@ucf.edu.cu, mciglesias@ucf.edu.cu

Italy and Spain was studied because they have passed the peak of the epidemic and because Cuba has used a confinement variant similar to that used by these countries.

As fundamental results, there are good model adjustments with respect to real values, for Italy and Spain. This allows confidence in the forecasts made for Cuba.

The predictions made for Cuba suggest that the estimated peaks of infections and deaths are reached between April 14 - 19 and April 18 - May 5, respectively. A total of infected between 1482 to 2678 and deaths between 63 to 211 is forecast.

Key words: Covid-19, predictive models, curve fitting, Cuba

1. INTRODUCCIÓN.

En la actualidad la humanidad presenta una crisis de salud debido al nuevo coronavirus SARS.CoV.2 causante de la enfermedad Covid-19. El virus es de fácil contagio, rápida propagación y presenta alto por ciento de mortalidad comparado con otros tipos de coronavirus.

Muchos gobiernos se encuentran inmersos en el despliegue de medidas para evitar el contagio y la muerte de personas en los respectivos países, aplicando en la mayoría de los casos, alguna variante de confinamiento, puesto que, realmente ha sido la medida que ha frenado en algunos países la propagación de la enfermedad. Es de vital importancia para estos gobiernos conocer un aproximado de la cantidad máxima de contagios y decesos, así como el momento de mayores casos para poder tomar medidas de tipo logístico.

En Cuba, al igual que en muchos otros países, se han puesto en marcha medidas para mitigar las afectaciones de esta enfermedad. La dirección del país a partir del 24 de marzo toma medidas orientadas a limitar la entrada de extranjeros al país, limitar el movimiento dentro de cada provincia y entre estas, evitar concentraciones de personas, cerrar temporalmente centros culturales y educativos, realizar pesquisas para encontrar personas con síntomas de la enfermedad y realizar pruebas a casos sospechosos para confirmar la enfermedad. Sin embargo, como ha sucedido en la mayoría de países, se ha manifestado un crecimiento aproximadamente exponencial, sobre todo de casos confirmados. Por ello es necesario contar con estimaciones que puedan servir para la toma de decisiones con los recursos disponibles.

Muchos de estos pronósticos se obtienen a partir de modelaciones matemáticas. Una modelación clásica que ha sido aplicada a las epidemias consiste en los modelos SIR basados en ecuaciones diferenciales ordinarias. Esta modelación ha sido utilizada con éxitos en la epidemia provocada por la Covid-19 en algunas regiones.[1–7]

Otras técnicas utilizadas para la modelación de la Covid-19 han estado basadas en:

Modelos estadísticos de series cronológicas para predecir el número de casos infectados y/o fallecidos.[3]

Procesamiento de información para la obtención de modelos predictivos a través del uso de internet.[14]

Modelos basados en inteligencia artificial y Machine Learning.[4,16]

Las modelaciones antes expuestas tienen presente una serie de parámetros que permiten la inclusión de varios factores con el fin de expresar, lo mejor posible, las realidades de las epidemias, sin embargo, presentan alto grado de complejidad y nivel de procesamiento para la obtención de estos parámetros.

Dentro las modelaciones utilizando modelos estadísticos se encuentran el ajuste de modelos de crecimiento poblacional logísticos y el modelo de crecimiento de Gompertz. Estos modelos han sido utilizados con éxito en la epidemia de la Covid-19 y presentan una complejidad inferior a las modelaciones antes expuestas.[5,13] Para la obtención de los parámetros de estos modelos se utiliza el método de los mínimos cuadrados para el caso de modelos lineales (MCL) y no lineales (MCNL) con respecto a los parámetros.

La curva de crecimiento logístico utilizada para la modelación de crecimientos poblacionales puede ser utilizada para pronosticar el crecimiento de casos infectados y/o decesos.[11] Un modelo básico logístico tiene la siguiente expresión:

$$P(t) = \frac{1}{1 + e^{b-at}}$$

Ecuación 1. Modelo básico logístico.

Este modelo puede ser utilizado para el pronóstico de las epidemias teniendo presente que $P(t)$ representa los casos acumulados (casos confirmados o fallecimientos), *t* representa el tiempo trascurrido luego de haberse presentados los primeros casos y *a* y *b* son parámetros que pueden ser obtenido (luego de realizar transformaciones matemáticas) aplicando el método de los MCL. Para la utilización de este modelo los datos deben ser llevados a la escala de [0; 1] dividiendo cada dato entre el acumulado. En este modelo un valor importante lo representa el punto de inflexión debido a que muestra el cambio de comportamiento de la curva lo cual puede ser interpretado como el pico de la epidemia. Este punto de inflexión está dado por la expresión:

$$t = \frac{b}{a}$$

Ecuación 2. Punto de inflexión del modelo logístico.

Una modelación que ofrece mayor versatilidad, está representada en la Ecuación 3. En este caso no es necesaria la transformación de los datos a la escala de [0; 1], dado que el parámetro *c* representa el máximo de casos acumulados y el punto de inflexión es el mismo que el presentado anteriormente. Sin embargo, presenta un mayor nivel de complejidad para la obtención de los parámetros debido a que no es posible, mediante transformaciones matemáticas, obtener un modelo lineal con respecto a los parámetros por lo que se aplica el método de los MCNL para la obtención de los mismos.[11]

$$P(t) = \frac{c}{1 + e^{b-at}}$$

Ecuación 3. Modelo logístico.

La curva de crecimiento exponencial de Gompertz pertenecen a la familia de curvas sigmoideas, inicialmente cóncavas y tras pasar el punto de inflexión, convexas. Su nombre procede del matemático Benjamin Gompertz, el cual utiliza la curva para describir la ley de la naturaleza que rige la mortalidad humana. Existen diferentes tipos de curvas Gompertz en función de los parámetros que la componen, sin embargo, están caracterizadas por una doble exponencial como elemento característico común.[11]

$$G(t) = ae^{-be^{-ct}}$$

Ecuación 4. Modelo de crecimiento de Gompertz.

En este modelo $G(t)$ representa los casos acumulados (casos confirmados o fallecimientos), *t* representa el tiempo trascurrido luego de haberse presentados los primeros casos y *a, b* y *c* son parámetros que pueden ser obtenido aplicando el método de los MCNL. El punto de inflexión está dado por la expresión:

$$t = \frac{\ln(b)}{c}$$

Ecuación 5. Punto de inflexión del modelo de crecimiento de Gompertz.

Para la bondad del ajuste de los modelos se utiliza el coeficiente de determinación ajustado R^2, adecuado para modelos no lineales con respecto a los parámetros.

En la presente investigación se pretende realizar pronósticos de las fechas para el pico de casos confirmado y de fallecidos por la enfermedad Covid-19 para Cuba, así como el total de casos confirmados y fallecidos pronosticados. Para ello son contrastadas estas ecuaciones con los datos de casos contagiados y fallecidos de Italia y España respectivamente, países que han pasado el pico de la epidemia y que, de alguna manera, Cuba ha implementado variantes de confinamiento similar a la de estos.

Se infiere que los modelos presentan idoneidad con los datos reales de Italia y España y por tanto puede dar buenos pronósticos para Cuba.

2. MÉTODOS.

Los datos utilizados para la estimación de los modelos fueron tomados del sitio web [8] publicados por la universidad Johns Hopkins. Estos datos se encuentran actualizados hasta la fecha de escrito este informe, 22 de abril de 2020.

Para seleccionar el primer día de conteo de datos se utilizó la siguiente metodología: se comenzó a contabilizar el primer día de aparición de casos confirmados o casos fallecidos, y se registraron los casos acumulados en ambos análisis. En la tabla 1 se muestran estas fechas para cada uno de los países analizados.

Tabla 1. Día de inicio del conteo de casos confirmados y de contagiados por país.

Fechas de inicio de conteo	**Datos acumulados registrados**	
	Contagiados	**Muertes**
Italia	31 de enero 2020	21 de febrero 2020
España	1 de febrero 2020	3 de marzo 2020
Cuba	12 de marzo 2020	18 de marzo 2020

Para el procesamiento de la información se utilizaron los programas informáticos Maxima 5.41.0 [7] y R 3.6.1 [10]. Para el procesamiento de los modelos que utilizaron el método de los MCL se empleó el programa simbólico Maxima y para el procesamiento de los modelos utilizando el método de MCNL se empleó el lenguaje de programación R, de procesamiento numérico.

En ambos casos se implementaron una serie de instrucciones que permiten el análisis, de forma rápida, de cualquier país afectado por la Covid-19 existente en la base de datos. Ambos ficheros de instrucciones (covid_19.mac para el caso de Maxima y covid_19.R para el caso de R) pueden ser descargados desde la dirección.

Para la utilización del método de los mínimos cuadrados se empleó el paquete lsquare.mac en el programa Maxima y las instrucciones nls, SSlogis y SSgompertz del paquete stat de R.

Se procesaron varios modelos de un mismo tipo en dependencia de la cantidad de datos que se iban incorporando para estimar la cantidad de días necesarios que permitieron, en los casos de España e Italia, una estabilidad en los modelos para luego establecer una comparación con el caso Cuba.

Fueron utilizados para el contraste de los modelos con los datos reales los picos ya alcanzados por Italia (pasados 56 días luego de haber aparecido el primer caso) y España (pasados 60 días luego de haber aparecido el primer caso). Estos picos, para Italia, corresponden al 26 de marzo 2020 para casos confirmados y 27 de marzo 2020 para casos fallecidos.[12] En el caso de España los picos corresponden al 31 de marzo 2020 para el números de contagios y 2 de abril 2020 para los decesos.[9]

Hasta la fecha de redacción de este reporte de investigación Italia había reportado un total 187327 casos confirmados por la Covid-19 con 25085 decesos, mientras que España había registrado 208389 casos confirmados y 21717 fallecidos.

Se utilizó el coeficiente de determinación ajustado R^2, para estimar la bondad del ajuste de los modelos, teniendo presente que valores mayores o iguales a 0.90 fueron tomados como adecuados.

3. RESULTADOS.

Casos confirmados por la Covid-19.

ESTUDIO DE CASO ITALIA.

El primer caso que aparece en Italia se diagnostica el 31 de enero de 2020. Sin embargo, fue hasta el 21 de febrero que realmente comienza un crecimiento exponencial de esta epidemia en esa nación. La figura 1 muestra la representación geométrica de los casos confirmados acumulados y los modelos logístico (ecuación 3) y curva de crecimiento exponencial de Gompertz (ecuación 4). El modelo logístico presenta un R^2 adecuado aproximadamente igual a 0,99717 y mediante este se estima el pico pasado 60 días (30 de marzo) luego de que apareció el primer caso. El modelo de Gompertz presenta un R^2 muy adecuado aproximadamente igual a 0,99962 y se estima el pico pasado 57 días (27 de marzo) luego de aparecido el primer caso es este país (31 de enero).

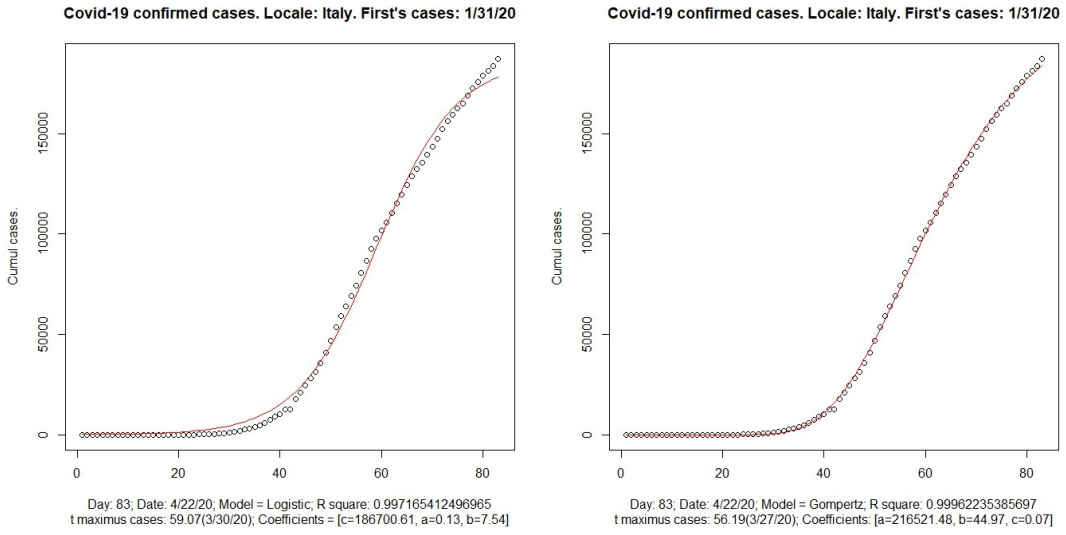


Figura 1. Modelo logístico y curva de crecimiento exponencial de Gompertz. Casos confirmados acumulados en Italia. (Tomado de R)

Procesamiento con mayor número de estos modelos en consideración en función de la cantidad de datos que se fueron incorporando en el transcurso de los días pusieron de manifiesto que se comenzaron a obtener pronósticos confiables (con $R^2 \geq 0.99$) pasado 38 días luego de que apareció el primer caso confirmado.

ESTUDIO DE CASO ESPAÑA.

En España el primer caso confirmado aparece el 1 de febrero de 2020. Sin embargo, fue hasta el 25 de febrero que realmente comienza un crecimiento exponencial de esta epidemia en el país. La figura 2 muestra la representación geométrica de los casos confirmados acumulados y los modelos logístico (ecuación 3) y curva de crecimiento exponencial de Gompertz (ecuación 4). El modelo logístico presenta un R^2 adecuado aproximadamente igual a 0,99747 y mediante este se estima el pico pasado 62 días (2 de abril) luego de que apareció el primer caso. El modelo de Gompertz presenta un R^2 muy adecuado aproximadamente igual a 0,99944 y se estima el pico pasado 59 días (30 de marzo) luego de aparecido el primer caso es este país (1 de febrero).

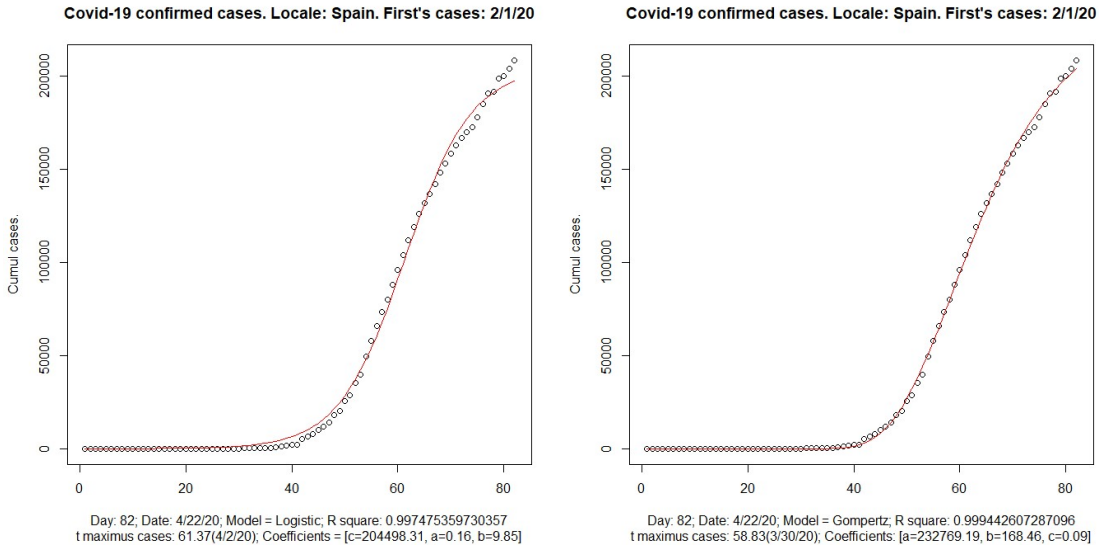


Figura 2. Modelo logístico y curva de crecimiento exponencial de Gompertz. Casos confirmados acumulados en España. (Tomado de R)

Procesamiento con mayor número de estos modelos en consideración en función de la cantidad de datos que se fueron incorporando en el transcurso de los días pusieron de manifiesto que se comenzaron a obtener pronósticos confiables (con $R^2 \geq 0.99$) pasado 40 días luego de que apareció el primer caso confirmado.

A continuación, se presenta una comparación de las estimaciones realizadas mediante los modelos para los casos contagiados de Italia y España (Tabla 2).

Tabla 2. Comparación de modelos estimados con los datos reales para los casos confirmados de la Covid-19 en Italia y España.

Estimación casos confirmados.		**Coeficientes (modelo)**	R^2	**Casos máximos pronosticados**	**Pico (pronóstico)**	**Pico real**
Italia	**Modelo logístico básico (Ecuación 1)**	a=0.21242 b=12.6156	0.96005	-	30 de marzo	26 de marzo (se registraron 6203 casos)
	Modelo logístico	a=0.13 b=7.54	0.99717	186701 (acumulado)	30 de marzo	

	(Ecuación 3)	c=186700.61				
	Modelo de Gompertz (Ecuación 4)	a=216521.48 b=44.97 c=0.07	0.99962	216522 (acumulado)	27 de marzo	
España	**Modelo logístico básico (Ecuación 1)**	a=0.2347 b=8.8653	0.96947	-	2 de abril	31 de marzo (se registraron 9222 casos)
	Modelo logístico (Ecuación 3)	a=0.16 b=9.85 c=204498.31	0.99748	204499 (acumulado)	2 de abril	
	Modelo de Gompertz (Ecuación 4)	a=232769.19 b=168.46 c=0.09	0.99944	232770 (acumulado)	30 de marzo	

Para ambos países el modelo de Gompertz presentó mejores estimaciones con R^2 muy adecuados y estimaciones muy buenas para los respectivos picos (con un día de error). El pronóstico de contagios oscila entre un rango de 186701 a 216522 para el caso de Italia y de 204499 a 232770 para el caso de España.

Casos fallecidos por la Covid-19.

ESTUDIO DE CASO ITALIA.

El primer fallecido en Italia se reportó el 21 de febrero. El gráfico mostrado en la figura 3 expone la representación geométrica de los casos fallecidos acumulados y los modelos logístico (ecuación 3) y curva de crecimiento exponencial de Gompertz (ecuación 4). El modelo logístico presenta un R^2 adecuado aproximadamente igual a 0,99690 y mediante este se estima el pico pasado 41 días (1 de abril) luego del reporte del primer fallecimiento. El modelo de Gompertz presenta un R^2 muy adecuado aproximadamente igual a 0,99964 y se estima el pico pasado 39 días (30 de marzo) luego del reporte del primer fallecimiento en este país.

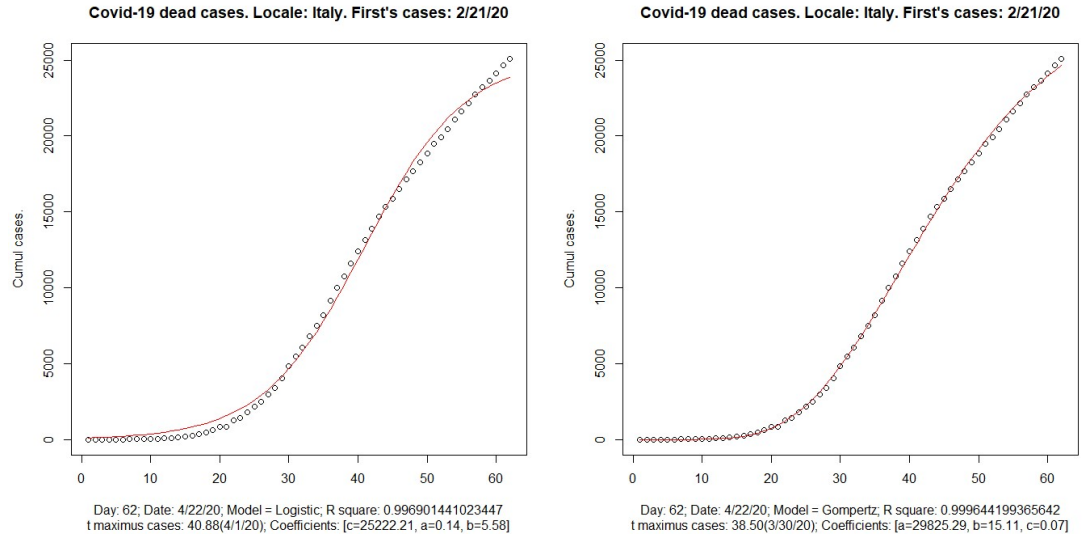


Figura 3. Modelo logístico y curva de crecimiento exponencial de Gompertz. Casos fallecidos acumulados en Italia. (Tomado de R)

Procesamiento con mayor número de estos modelos en consideración en función de la cantidad de datos que se fueron incorporando en el transcurso de los días pusieron de manifiesto que se comenzaron a obtener pronósticos confiables (con $R^2 \geq 0.99$) pasado 30 días luego de que se registró el primer caso fallecido.

ESTUDIO DE CASO ESPAÑA.

El primer fallecido en España se reportó el 3 de marzo. La figura 4 muestra la representación geométrica de los casos fallecidos acumulados y los modelos logístico (ecuación 3) y curva de crecimiento exponencial de Gompertz (ecuación 4). El modelo logístico presenta un R^2 adecuado aproximadamente igual a 0,99683 y mediante este se estima el pico pasado 33 días (4 de abril) luego del reporte del primer fallecimiento. El modelo de Gompertz presenta un R^2 muy adecuado aproximadamente igual a 0,99971 y se estima el pico pasado 30 días (1 de abril) luego del reporte del primer fallecimiento en este país.

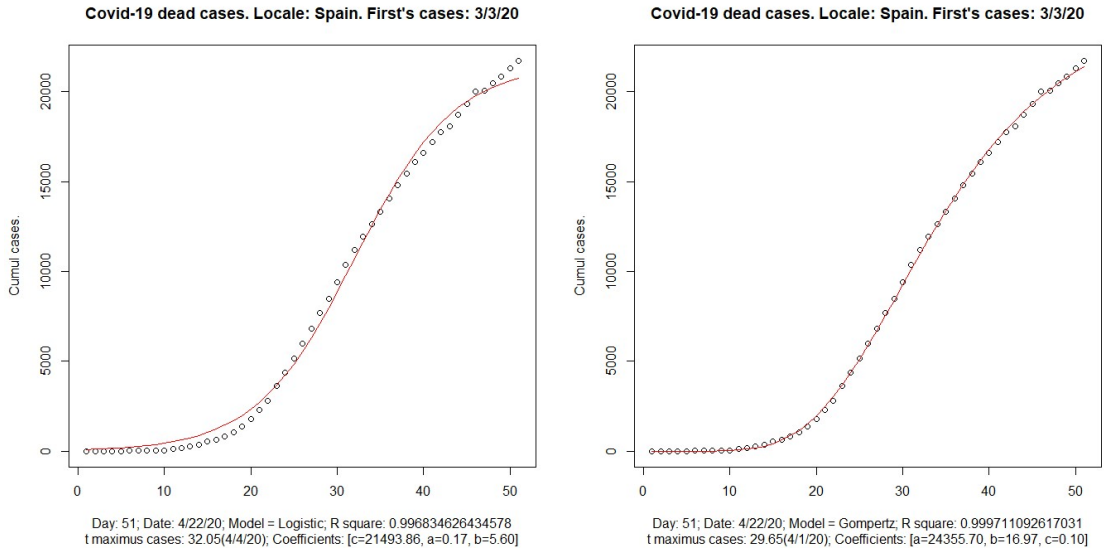


Figura 4. Modelo logístico y curva de crecimiento exponencial de Gompertz. Casos fallecidos acumulados en España. (Tomado de R)

Procesamiento con mayor número de estos modelos en consideración en función de la cantidad de datos que se fueron incorporando en el transcurso de los días pusieron de manifiesto que se comenzaron a obtener pronósticos confiables (con $R^2 \geq 0.99$) pasado 20 días luego de que se registró el primer caso fallecido.

A continuación, se presenta una comparación de las estimaciones realizadas mediante los modelos para los casos fallecidos de Italia y España (Tabla 3).

Tabla 3. Comparación de modelos estimados con los datos reales para los casos fallecidos por la Covid-19 en Italia y España.

Estimación casos fallecidos.		**Coeficientes (modelo)**	R^2	**Casos máximos pronosticados**	**Pico (pronóstico)**	**Pico real**
Italia	**Modelo logístico básico (Ecuación 1)**	a=0.20465 b=8.3202	0.96876	-	1 de abril	27 de marzo (se registraron 919 casos)
	Modelo logístico (Ecuación 3)	a=0.14 b=5.58 c=25222.21	0.99690	25223 (acumulado)	1 de abril	
	Modelo de Gompertz (Ecuación 4)	a=29825.29 b=15.11 c=0.07	0.99964	29826 (acumulado)	30 de marzo	
España	**Modelo logístico básico (Ecuación 1)**	a=0.2555 b=8.3275	0.96943	-	3 de abril	2 de abril (se registraron 961 casos)
	Modelo logístico (Ecuación 3)	a=0.17 b=5.60 c=21493.86	0.99683	21494 (acumulado)	4 de abril	
	Modelo de Gompertz (Ecuación 4)	a=24355.70 b=16.97 c=0.10	0.99971	24356 (acumulado)	1 de abril	

Para ambos países el modelo de Gompertz presentó mejores estimaciones con R^2 muy adecuados y estimaciones muy buenas para ambos picos (con tres días de error para Italia y un día para el caso de España). El pronóstico de fallecidos oscila entre un total de 25223 a 29826 para Italia y de 21494 a 24356 para España.

Estimación para el caso de Cuba.

Cuba presentó sus primeros casos de contagios el 11 de marzo, pero se registraron a partir del día posterior (12 de marzo) y el primer deceso el 18 de marzo. Al momento de escrito este informe habían transcurrido 42 días desde el primer reporte de contagios y 36 días desde el primer fallecimiento. La figura 5 expone la representación geométrica de los casos confirmados y fallecidos acumulados y los modelos: logístico (ecuación 3) y curva de crecimiento exponencial de Gompertz (ecuación 4).

Para los casos confirmados el modelo logístico presenta un R^2 adecuado aproximadamente igual a 0,99865 y mediante este se estima el pico pasado 34 días (14 de abril) luego de que apareció el primer caso. El modelo de Gompertz presenta un R^2 muy adecuado aproximadamente igual a 0,99936 y se estima el pico pasado 39 días (19 de abril) luego de aparecido el primer caso es este país.

Para los casos fallecidos el modelo logístico presenta un R^2 adecuado aproximadamente igual a 0,99656 y mediante este se estima el pico pasado 32 días (18 de abril) luego del reporte del primer fallecimiento en el país. El modelo de Gompertz presenta un R^2 adecuado aproximadamente igual a 0,99582 y se estima el pico pasado 49 días (5 de mayo) luego de aparecido el primer caso en el país.

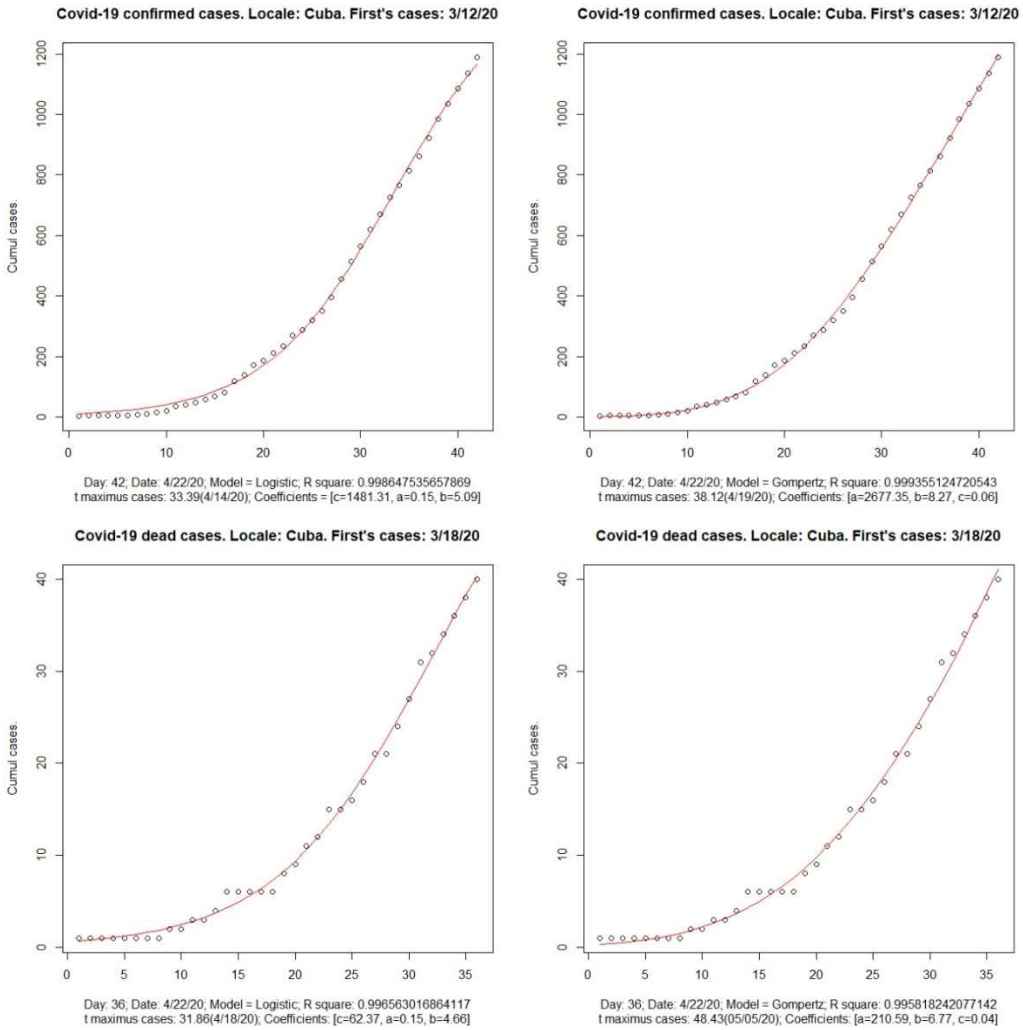


Figura 5. Modelo logístico y curva de crecimiento exponencial de Gompertz. Casos confirmados y fallecidos acumulados en Cuba. (Tomado de R)

4. DISCUSIÓN.

Los modelos de crecimiento logístico (ecuación 3) y de Gompertz (ecuación 4) dieron buenos pronósticos para el caso de Italia y España. Para el caso de Italia el modelo de Gompertz dio pronósticos con error de 1 posterior y 3 días posteriores para el pico de casos contagiados y fallecidos respectivamente con respecto a los picos reales presentados por este país. En el caso de España el modelo de Gompertz presentó pronósticos de los picos de contagiados y fallecidos con un día de error anterior a los días reales en que se produjeron dichos picos.

Estos pronósticos han sido confiables para procesamiento de datos cercanos a los 40 días luego de haberse presentado los primeros casos confirmados y 30 días pasado el primer reporte de deceso en ambos países. Italia presentó su pico de casos contagiados 56 días luego del reporte del primer contagio y España 60 días, sin embargo, ambos países comenzaron como tal a tomar medias serias alrededor de 20 días después del primer contagio, cuando se presentaron los primeros fallecimientos. De igual forma Italia presentó su pico de casos fallecidos 36 días luego del primer deceso y España 31 días con posterioridad al primer reporte de fallecimiento.

Cuba, en el momento de escrito este informe, se encuentra pasado 42 días luego del primer reporte de casos positivos a la enfermedad y 36 días pasados del primer registro de deceso. Esto hace que los datos, a partir de los resultados obtenidos para España e Italia, sea adecuados para obtener modelos con buenos pronósticos. Sin embargo, se debe señalar que los pronósticos arrojados por los modelos dan el pico estimado de contagiados en Cuba pasado entre 34 y 39 días luego de haberse confirmado los primeros casos positivos a la Covid-19 en el país (12 de marzo) y el pico de fallecidos pasado entre 32 a 49 días luego de haberse confirmado el primer deceso en el país (18 de marzo).

Este comportamiento arrojado por los modelos de pronóstico se invierte con respecto a las estadísticas mostradas por los dos países europeos analizados, que presentaron en menor tiempo, a partir del primer caso contabilizado, el pico de casos fallecidos con respecto al de casos contagiados (aunque cronológicamente en ambos países el pico de contagiados fue antes que el pico de fallecidos).

Aunque Italia y España han sido seleccionadas para contrastar los modelos debido a que ya pasaron su pico y a las similitudes con respecto a las estrategias de confinamiento utilizadas por ellos y Cuba, es bueno destacar que existen importantes diferencias que pudieran introducir sesgos en las estimaciones realizadas para el caso de Cuba.

Una primera, tangible e importante diferencia viene dada en la cantidad de datos que fueron utilizados para el procesamiento y estimación de los parámetros de los modelos. El caso de Cuba, presenta un aproximado de 20 datos menos.

Otras diferencias vienen dadas por que Cuba presenta una menor población, con menor cantidad de ciudades importantes (o grandes ciudades), menor movimiento interno y menor cantidad de visitantes extranjeros, que Italia y España.

Cuba comienza a ser afectada por la enfermedad Covid-19 alrededor de 20 días después de haber sido afectados estos dos países, lo cual le dio un margen de preparación basado en las experiencias de países como China, Corea del Sur y los propios Italia y España.

El contagio llegó a ser descontrolado en Italia y España, sin embargo, esa característica, según medios oficiales, no se ha presentado a fecha de escrito este informe en Cuba.

El sistema de salud cubano es diferente al que presentan estas dos naciones y la política por parte del sistema de salud y el gobierno para tratar la posible epidemia es también distinta.

5. CONCLUSIONES.

En el presente artículo de investigación han sido utilizados modelos logísticos y de crecimiento de Gompertz para obtener pronósticos de los días de mayor afectación para los casos contagiados y fallecidos por la Covid-19, así como la cantidad total de los mismos en Cuba.

Estos modelos han sido contrastados, con resultados satisfactorios, con los datos publicados sobre las afectaciones de esta enfermedad para los países de Italia y España quienes han pasado el pico de la epidemia y utilizaron políticas de confinamientos similares a las que lleva a cabo Cuba.

El pico estimado de contagiados se debe haber alcanzado (con respecto a la fecha en que se ha escrito este informe) entre el 14 y 19 de abril, pronosticándose una cantidad total de contagios entre 1482 a 2678. Sin embargo, el pico de fallecidos se pronostica que debe alcanzarse entre el 18 de abril y 5 de mayo y se estima una cantidad total de deseos en un rango entre 63 a 211.

REFERENCIAS

1. Bacaër N: Un modelo matemático de la epidemia de coronavirus en Francia. no date; .
2. Cagigal MAG, Becario FPU: Modelado y análisis de la evolución de una epidemia vírica mediante filtros de Kalman: el caso del COVID-19 en España. no date; .
3. Deb S, Majumdar M: A time series method to analyze incidence pattern and estimate reproduction number of covid-19. ArXiv Prepr ArXiv200310655. 2020; .
4. Hu Z, Ge Q, Jin L, Xiong M: Artificial intelligence forecasting of covid-19 in china. ArXiv Prepr ArXiv200207112. 2020.
5. Jia L, Li K, Jiang Y, Guo X: Prediction and analysis of Coronavirus Disease 2019. ArXiv Prepr ArXiv200305447. 2020.
6. Li C, Chen LJ, Chen X, Zhang M, Pang CP, Chen H: Retrospective analysis of the possibility of predicting the COVID-19 outbreak from Internet searches and social media data, China, 2020. Eurosurveillance. 2020; 25:2000199.
7. Maxima, a Computer Algebra System: no date; .
8. Novel Coronavirus (COVID-19) Cases Data - Humanitarian Data Exchange: no date;
9. RTVE.es: Curva de casos y muertes por Coronavirus en España, día a día. RTVE.es. 2020; .
10. R: The R Project for Statistical Computing: 2018; .
11. Simón Mínguez F: Procesos de difusión Logístico y Gompertz. Métodos numéricos clásicos en la estimación paramétrica. 2016.
12. Tiempo CEE: Italia llegó al pico de contagios, según Instituto de Sanidad. El Tiempo. 2020.
13. Villalobos-Arias M: Estimation of population infected by Covid-19 using regression Generalized logistics and optimization heuristics. ArXiv Prepr ArXiv200401207. 2020; .
14. Wang CJ, Ng CY, Brook RH: Response to COVID-19 in Taiwan: big data analytics, new technology, and proactive testing. Jama. 2020;
15. Yang Z, Zeng Z, Wang K, et al.: Modified SEIR and AI prediction of the epidemics trend of COVID-19 in China under public health interventions. J Thorac Dis. 2020; 12:165.
16. Zhou C, Su F, Pei T, et al.: COVID-19: challenges to GIS with big data. Geogr Sustain. 2020; .

Capítulo 2 pp 13-29

NUEVO ESCENARIO DE VULNERABILIDAD Y POBREZA A CAUSA DE COVID-19. PLANEACIÓN ESTRATÉGICA PARA EL IMPULSO DEL DESARROLLO LOCAL EN MÉXICO.

García Rodríguez José Félix[1], Hernández Govea Luis Manuel[1], Caamal Cauich Ignacio[2], Pineda Celaya Lourdes del Carmen[3], Izquierdo Balcázar Naamán[1]

[1]Universidad Juárez Autónoma de Tabasco. UJAT
[2]Universidad Autónoma de Chapingo. UACH
[3]Universidad Popular Autónoma del Estado de Puebla

RESUMEN

La emergencia sanitaria global causada por la pandemia de coronavirus SARS-CoV-2 que provoca la enfermedad COVID-19 ha impactado negativamente el bienestar de la población y conduce a todos a replantear las prioridades individuales y colectivas. En México, a pesar ser un país de ingreso medio, están presentes inaceptables niveles de pobreza y rezago socioeconómico, y los efectos de la pandemia hacen más evidente dicha situación. El estado mexicano cuenta con instrumentos de planeación estratégica para enfrentar esta problemática. El **objetivo** de esta investigación es proponer un plan estratégico para el desarrollo local de las comunidades en situación de pobreza, tomando como referente una localidad tipo. **Hipótesis.** Si se aplicara un plan estratégico para el desarrollo local enfocado a las comunidades rurales en situación de pobreza, entonces se identificarían los factores determinantes y condicionantes del problema, se establecerían los medios adecuados para enfrentarlo, y se focalizarían eficientemente las políticas públicas establecidas. **Método.** Investigación social con enfoque cuantitativo de tipo descriptivo y transversal, aplicada en una comunidad en situación de pobreza. De un universo poblacional de 500 familias, y mediante muestreo aleatorio simple con intervalo de confianza del 95% y margen de error del 5%, se determinó una muestra de 90 hogares. Se aplicó un cuestionario estructurado a partir de investigaciones precedentes. **Resultados.** Se identificaron los factores determinantes de pobreza, así como sus fortalezas y debilidades, con lo que se elaboró un plan estratégico prototipo para el desarrollo local.

PALABRAS CLAVE: Planeación estratégica, Desarrollo, Desarrollo local, Pobreza, Plan estratégico.

ABSTRACT

In Mexico, the State is responsible for national economic planning for the development and well-being of the population, especially those facing problems of poverty and social backwardness. For this, it has strategic planning as a methodological analysis tool. The **objective** of the research is to draw up a strategic plan for the local development of communities in poverty, taking as a reference a type locality. **Hypothesis.** If a strategic

plan for local development focused on rural communities in poverty were applied, then the determining and conditioning factors of the problem would be identified, the appropriate means to deal with it would be established, and the established public policies would be efficiently targeted. **Method.** Social research with a descriptive and transversal quantitative approach, applied in a community in a situation of poverty. From a population universe of 500 families, and by simple random sampling with a 95% confidence interval and a 5% margin of error, a sample of 90 households was determined. A structured questionnaire was applied based on previous research. **Results**. The determinants of poverty were identified as well as their strengths and weaknesses, with which a prototype strategic plan for local development was prepared.

KEY WORDS: Strategic planning, Development, Local development, Poverty, Strategic plan.

1. INTRODUCCIÓN

Constitucionalmente, en México el Estado es responsable de la planeación económica nacional que impulse el desarrollo y bienestar de la población. En materia de desarrollo se ha avanzado mucho en el plano regional, sobre todo en el norte del país, donde algunos estados cuentan con alto grado de desarrollo económico e industrial y mejores condiciones de vida de sus habitantes, sobre todo en sus áreas urbanas. No obstante, en el sur-sureste mexicano prevalece el rezago socioeconómico e industrial. Por ello, existe necesidad de impulsar el desarrollo homogéneo del país, así como el bienestar económico y social en el ámbito local, disminuyendo los niveles de pobreza y desigualdad, rezago e inseguridad presentes en buena parte del país (Miguel (2004).

La pobreza en México es un problema socioeconómico de naturaleza compleja y multifactorial, que está presente en las pequeñas y grandes urbes sociales, así como en el ámbito rural. Sin embargo, son los habitantes de las comunidades y pueblos rurales quienes sufren más sus consecuencias: falta de empleo e ingresos, carencias sociales, rezago educativo, falta de acceso a la alimentación, desnutrición, inseguridad, etc., factores característicos del subdesarrollo y rezago social que los obliga a migrar de sus lugares de origen. Ello a pesar del derecho fundamental a una vida digna. Ante ello, es responsabilidad del Estado nacional asegurar el cumplimiento del mismo, creando leyes, políticas y programas que impulsen el bienestar de la población, especialmente de los más vulnerables.

Para lograr este propósito, el Estado cuenta con una importante herramienta metodológica de apoyo como lo es la planeación económica nacional para el desarrollo regional y local. En este contexto, la necesidad de encausar el desarrollo de las comunidades en situación de pobreza, podría ser atendida a través de la integración de una propuesta de plan estratégico metodológicamente aplicable a todas las comunidades que cumplan estas características. Dicho plan, cumpliéndose ciertas condiciones, podría ser replicado a otras localidades de características similares, de ahí su importancia metodológica. Con el propósito de aportar a esta solución, esta investigación tiene como **objetivo general**: Proponer un plan estratégico metodológico para impulsar el desarrollo local de las comunidades en situación de pobreza de México, así como contribuir al combate a la

pobreza y al desarrollo regional. Dicho plan contendrá los elementos metodológicos de la planeación estratégica que permitirán conocer las carencias y necesidades a nivel local, así como las potencialidades disponibles para impulsar el desarrollo local. Cabe mencionar que la propuesta de plan estratégico para el desarrollo local a integrar tiene como referente la investigación efectuada en la Ranchería El Golpe 2ª Sección del municipio de Cárdenas, en el estado de Tabasco, México. De la misma manera, la investigación es guiada por la siguiente **hipótesis**: Si se aplicara un plan estratégico de desarrollo local enfocado a las comunidades rurales de México, entonces se identificarían adecuadamente las condiciones de pobreza, se determinarían los medios más óptimos para combatirla y se focalizarían mejor las políticas de impulso al desarrollo local. Para lograrlo, se desarrolló una **investigación cuantitativa** aplicada a una localidad tipo del municipio de Cárdenas, Tabasco, misma que cumple las características requeridas para la investigación, como son alta pobreza y marginación. El diseño contempló un universo poblacional de 116 hogares en los que viven los 500 habitantes de la localidad. El tamaño de muestra es de 90 hogares, y se determinó mediante muestreo aleatorio simple, considerándose un nivel de confianza del 95% y un margen de error del 5%.

La investigación es relevante, toda vez que aporta una propuesta metodológica para el desarrollo local de las comunidades en situación de pobreza. El Plan Estratégico de Desarrollo Local, producto de la investigación podría constituirse en un instrumento de trabajo para la planificación pública del desarrollo local en los tres niveles de gobierno. Además, constituiría una importante herramienta metodológica al alcance de las comunidades rurales mediante la cual podrán involucrarse en su propio proceso de cambio, identificando sus potencialidades, fortalezas y debilidades, así como las organizaciones a constituir y las inversiones a realizar para encausar el desarrollo local, haciéndolos de esta manera sujetos de su propio desarrollo. De esta manera, se espera que la investigación sea un referente en la gestión del desarrollo local integral.

Enseguida se presentan los principales elementos teóricos que dan sustento a la investigación, los cuales al concluir el apartado se sistematizan en un algoritmo teórico sobre el proceso de desarrollo local, mismo que orienta metodológicamente el diseño del estudio, así como la propuesta planteada.

1.1. Desarrollo y pobreza

Desarrollo. La conceptualización del desarrollo surge a partir de la posguerra como un tema de análisis y reflexión, siendo en la Carta del Atlántico, firmada en 1941 por Churchill y Roosevelt, donde se hace referencia a la seguridad económica y social universal como objetivo fundamental para garantizar la paz (Boisier, 2001). No obstante, el concepto cobra fuerza durante la presidencia de Truman, quien el 20 de enero de 1949 da indicaciones para la creación de diversos programas de apoyo para el desarrollo de los países pobres, cuya organización y supervisión recayó en el Banco Mundial (BM) y el Fondo Monetario Internacional (FMI) como organizaciones de apoyo a las Naciones Unidas (Miguel, 2004).

En un principio, el concepto de desarrollo fue asociado al de crecimiento económico, ya que los primeros teóricos entendían que la mejora de una sociedad se daría en la medida en que se incrementaran las inversiones y la productividad, lo que se vería reflejado en los ingresos de las familias y por consiguiente en la mejora de sus condiciones de vida. De esta manera, el Producto Interno Bruto (PIB) y el PIB per cápita son los indicadores macroeconómicos por excelencia que toman preponderancia (Boisier 2001). Sin embargo, surge de inmediato la pregunta obligada: ¿cuánta congruencia habrá entre el PIB per cápita y las condiciones reales de bienestar de las familias? Evidentemente, la debilidad de este indicador es que no mide la brecha de desigualdades sociales. Así, en la conceptualización del desarrollo debe estar implícita la mejora de las condiciones de vida de la sociedad, más allá del incremento de los ingresos. Por lo tanto, el desarrollo no es un hecho terminado en el tiempo, sino más bien un proceso en el que los objetivos y las acciones están encaminadas a garantizar el bienestar de la población de manera creciente y constante (Miguel, 2004).

Desarrollo económico. El desarrollo económico debe asumirse como un proceso a través del cual una nación logra alcanzar mejores estándares de vida de su población (Brue & Grant, 2009). En este contexto, la ONU (2012) ha planteado diversas prioridades para el logro de un desarrollo sostenible que fomente la prosperidad, las oportunidades económicas, el bienestar social y la protección del medio ambiente; todo ello con la intención de mejorar las condiciones de vida de la población en general. De esta manera, el desarrollo económico busca un equilibrio en sus tres pilares básicos: económico, social y ecológico. En este contexto, la pregunta central sería cómo lograr el desarrollo económico que impulse el bienestar de la sociedad, el crecimiento económico sostenido y el cuidado del medio ambiente. Los economistas coinciden en cuatro elementos indispensables para el desarrollo: recursos humanos, recursos naturales, capital e innovación y cambio tecnológico (Samuelson & Nordhaus, 2010).

Si bien es cierto que el propósito central del desarrollo económico es mejorar los niveles de vida de las personas, no debe dejarse de lado que no hay desarrollo sin crecimiento económico; es decir, el crecimiento económico está implícito en el desarrollo (Samuelson & Nordhaus, 2010). Así, aunque la teoría del desarrollo económico sostiene que el PIB per cápita no es un indicador objetivo para medir el bienestar de la población, Krugman y Wells (2014) explican que éste se usa como medida resumida del progreso del país a lo largo del tiempo, y que el crecimiento económico a largo plazo depende casi totalmente de la variable productividad, la cual obedece el aumento del capital físico, el aumento del capital humano y el progreso tecnológico.

En síntesis, no hay desarrollo económico sin crecimiento económico, pero el crecimiento en sí mismo no genera desarrollo económico. Esto justifica la intervención del Estado en la actividad económica a efectos de corregir las fallas del mercado, procurando una distribución justa y equitativa de la riqueza (Espinoza, 2008). No obstante, la historia económica demuestra que el crecimiento económico ha generado una gran brecha de desigualdad, misma que se traduce en marginación y pobreza. Esto se debe a que el retorno del capital es más alto que la renta del trabajo. Si bien el crecimiento económico no resuelve esta brecha social, no deja de ser un elemento fundamental del desarrollo económico (Piketty, 2014).

Desarrollo Regional. El desarrollo regional consiste en un proceso de cambio estructural localizado espacialmente, asociado a un proceso permanente de progreso de la propia región, así como de sus habitantes colectivamente e individualmente. Es decir, el desarrollo regional debe combinar tres dimensiones: la dimensión espacial, la dimensión social y la dimensión individual. Se trata de un proceso en el que la región es un sujeto colectivo (Boisier, 2001). El concepto de desarrollo regional se ha considerado en dos vertientes: subordinado al contexto nacional o independiente del mismo. En el primer caso, el desarrollo regional se ha entendido como un proceso de desarrollo nacional a escala regional, considerándose las características económicas, sociales y físicas del cambio en una zona durante un periodo de tiempo; en el segundo, el desarrollo regional se concibe como un aumento del bienestar en la región expresado en indicadores tales como el ingreso per cápita, su distribución entre los habitantes, el acceso a los servicios sociales y la adecuación de normas legales y administrativas. Aunque algunos prefieren ver el desarrollo regional como un proceso dependiente del desarrollo nacional, en muchas regiones pobres se han superado aspectos desfavorables o se han creado nuevas situaciones favorables que mejoran la calidad de vida de sus habitantes gracias a la planificación del desarrollo regional (Miguel , 2004).

Desarrollo Local. El desarrollo local es un proceso en el que una sociedad, a partir de su identidad y su propio territorio, genera y fortalece sus dinámicas económicas, sociales y culturales, facilitando la articulación de cada uno de estos subsistemas y logrando mayor intervención y control entre ellos. (Casanova, citado por Alcañiz, 2008). De esta manera, el desarrollo local procura determinar dos aspectos: primero, cuál es el potencial de recursos con el que se cuenta; y segundo, cuáles son las necesidades que se requiere satisfacer de las personas, comunidades, colectividades, municipios y la sociedad en general. El desarrollo local parte del análisis de los recursos disponibles con los que se impulsará el bienestar de una comunidad pequeña o grande, con base en un proceso de planeación estratégica (Silva & Sandoval, 2012). Este enfoque surge como respuesta a las fuertes transformaciones producidas por la acumulación de capital, que plantea problemas de regulación como la gestión de trabajo o la adaptación y difusión de la tecnología moderna que las instituciones del pasado afrontaron, pero que ahora son más complejas, de manera que los instrumentos de intervención estatal han perdido eficacia en la regulación de la economía, por lo que estos cambios obligan al Estado a intervenir estratégicamente. La reestructuración del Estado está impulsando formas nuevas en la gestión pública como es la política del desarrollo local. Ante esta problemática, muchas comunidades, especialmente europeas, han intentado dinamizar el ajuste de los sistemas productivos locales (Boisier, 2001).

Desarrollo Sustentable. El enfoque de desarrollo sustentable tiene sus orígenes en la década de los ochenta, cuando la ONU crea la Comisión Sobre el Medio Ambiente y el Desarrollo y su famoso informe Nuestro Futuro Común, también conocido como El Informe Brundtland. En el mismo se señalaba con claridad que la sociedad debía modificar su estilo y hábitos de vida para evitar una crisis social y la degradación de la naturaleza, de manera tal que el desarrollo sustentable debía satisfacer las necesidades de la generación presente, sin comprometer la capacidad de las generaciones futuras para satisfacer sus propias necesidades (Ramírez,

Sánchez, & García, 2004). La definición anterior contempla dos conceptos importantes: 1) el concepto de necesidades, especialmente las necesidades básicas de las personas vulnerables, y 2) preservación del medio, es decir, reconoce que las capacidades para la satisfacción de necesidades provienen de la naturaleza, por lo tanto, es responsabilidad de la generación presente satisfacer sus necesidades, sin comprometer las capacidades de satisfacción de las generaciones futuras.

En este contexto, en septiembre de 2015 se estableció la Agenda 2030 para el Desarrollo Sostenible, misma que contempla diecisiete objetivos para alcanzar el desarrollo sostenible, entre los que destaca el fin de la pobreza (ONU, 2015). Estos objetivos, como señala Zarta (2018) persiguen un proceso armonioso de desarrollo sustentable en el cual colaboren todas las disciplinas del conocimiento, especialmente en lo económico, social, ambiental, cultura, así como un sistema de valores correspondientes.

Pobreza. Uno de los problemas más debatidos dentro del ámbito de la economía, la política, la filosofía y la ética es la pobreza, entendida como una condición socioeconómica que limita el bienestar de las personas y que constituye un asunto de naturaleza multidimensional y complejo (García, 2016). Al tratarse de un problema complejo, su estudio y abordaje requiere enfoques multidimensionales e interdisciplinarios para comprender sus causas y diseñar políticas adecuadas para enfrentarlo. La pobreza es un problema latente a nivel mundial. Según el Banco Mundial (2019), para el 2018 había 7 mil 594 millones de habitantes en el mundo, de los cuales 736 millones se encontraban en situación de pobreza extrema. Esto quiere decir que el 10% de la población mundial se encuentra en condiciones de pobreza; personas que viven diariamente con un ingreso por debajo de 1.90 dólares.

En México la pobreza es un tema preocupante. Así, en el año 2018 de una población total de casi 125 millones de habitantes, 52.4 millones, se encontraban en situación de pobreza multidimensional, lo que equivale al 41.9% del total (CONEVAL, 2018). Esto implica que tal población tiene al menos una carencia social y su ingreso es insuficiente para adquirir los bienes y servicios que requiere para satisfacer sus necesidades alimentarias. Lo más alarmante es que el 7.4% de la población total se encontraba en pobreza extrema; es decir, 9.3 millones de personas vivían con un ingreso tan bajo que ni aun gastándolo exclusivamente en alimentos podían adquirir los nutrientes necesarios para una vida sana, y además tenían tres o más carencias sociales. Parte del problema de la pobreza radica en la baja efectividad de las políticas públicas implementadas para enfrentarla. Así, En el periodo 2015 al 2017 operaron en el país 5,491 programas sociales de los cuales sólo 83 estaban dirigidos a combatir las carencias sociales y eran considerados prioritarios. Sin embargo, de acuerdo al CONEVAL estos programas no dieron los resultados esperados por deficiencias en su diseño e implementación (Roldán, 2017).

Geográficamente, la mayor parte de la población en situación de pobreza se concentra en las zonas rurales, lo que obedece a la exclusión social y económica a la que están expuestas, y que se traduce en limitaciones de acceso a los servicios de educación y salud, así como a los mercados laborales, limitados canales de comercialización de los productos locales, así como deficientes vías de comunicación. Ante ello es necesario

promover la inclusión social y el desarrollo de capacidades locales (Portales, 2014). Para lograrlo, la planeación estratégica del desarrollo local se presenta como una alternativa efectiva para enfrentar la pobreza (Herrera, 2013). A partir del diseño de un plan estratégico de desarrollo local, el gobierno puede focalizar mejor las políticas y programas de apoyo a las comunidades, proveyéndoles mejores servicios básicos y mayores elementos para el desarrollo local sustentable.

1.2. Planeación estratégica

La planeación es la primera fase del proceso administrativo de cualquier organización, pública o privada. Para el Estado, la planeación es un instrumento de la política de desarrollo que hace posible establecer políticas efectivas de crecimiento, y facilita la toma de decisiones sobre bases objetivas y estratégicas (Chapoy, 2003). Según Chandler (2003), la estrategia consiste en el establecimiento de metas y objetivos a largo plazo de una organización, así como las acciones a emprender y los recursos necesarios para su consecución. Al respecto, Armijo (2011) comenta que la planeación estratégica es un ejercicio de formulación y establecimiento de objetivos prioritarios, cuya característica central es el establecimiento de los cursos de acción (estrategias) para alcanzarlos. La planeación estratégica está integrada por elementos consecutivos que dan orden a su elaboración: misión, visión, valores organizacionales, objetivos estratégicos, análisis interno y externo, estrategias, líneas de acción e indicadores de desempeño. Todos estos elementos se plasman en un documento rector denominado plan estratégico de desarrollo. En este contexto, la planeación del desarrollo local debe ser estratégica, a efectos de priorizar los objetivos que se propone alcanzar.

La planeación estratégica del desarrollo local exige la participación activa de los agentes involucrados, como son la población beneficiaria, autoridades locales, instituciones públicas y universidades, agencias de desarrollo local, organizaciones sociales, etc. (Silva & Sandoval, 2012). En general, la planeación estratégica local implica realizar un diagnóstico de las fortalezas y debilidades presentes en la localidad, y a partir de ello elaborar un plan estratégico que contenga las acciones que deben emprenderse para enfrentar los problemas sociales y económicos complejos que enfrentan, como son entre otros la pobreza, desigualdad y marginación.
En base a los referentes teórico-metodológicos antes expuestos, se determinó el siguiente algoritmo teórico para el desarrollo local, adaptado de la propuesta de Miguel (2004), mismo que en lo sucesivo orienta el diseño de la investigación y la propuesta planteada.

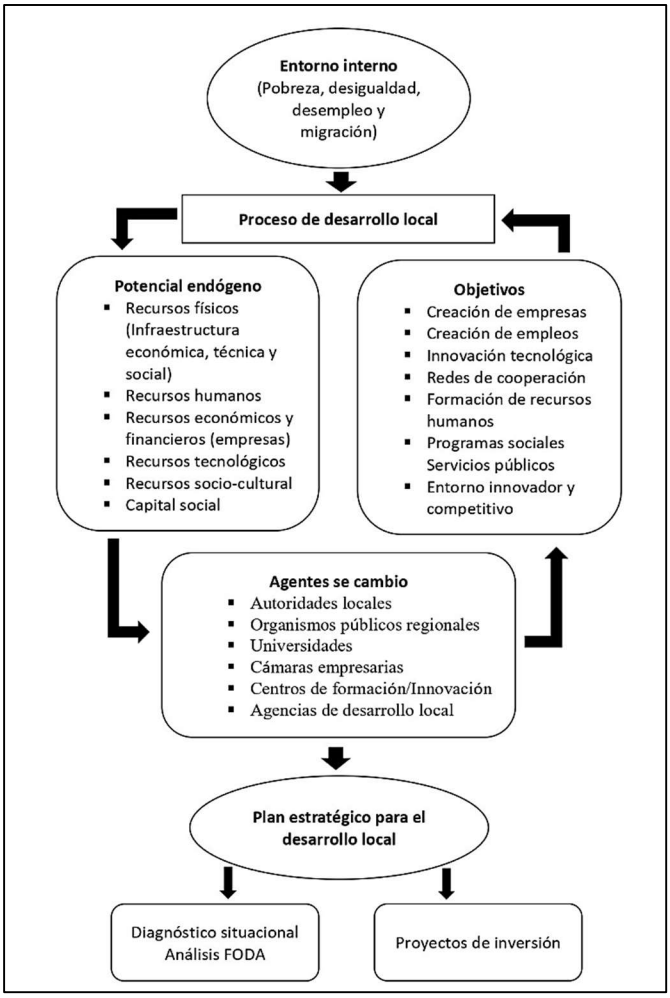


Figura 1: Algoritmo de la ruta del desarrollo local sustentable. (Adaptado de Miguel, 2004)

2. MÉTODO

Enfoque. Se aplica el enfoque cuantitativo confirmatorio**.** En éste se parte de una idea que deriva en el planteamiento de un problema de investigación, preguntas y objetivos; se sustenta en una revisión de la literatura actualizada y se establece un marco teórico, así como una hipótesis a demostrar, determinándose las variables de estudio, mismas que a partir de la información captada en el trabajo de campo son medidas en base a estadística descriptiva (Hueso y Cascant 2012). De esta manera, la investigación recoge, procesa y analiza los datos sobre las variables previamente determinadas y se estudia la relación entre estas, para posteriormente realizar la interpretación de los resultados (Surday, 2007). De la misma manera, en esta investigación social se aplica el método deductivo, ya que parte de lo general a lo particular. Esto es, se entiende el problema como un todo y luego se ocupa de la parte en la que se desea profundizar (Abreu 2014).

Tipo de investigación. Investigación descriptiva y transversal, cuyo propósito es describir y analizar las variables identificadas, así como la interrelación existente entre ellas (Hernández, Fernández, & Baptista, 2014).

Diseño. Como caso de estudio se seleccionó la localidad denominada Ranchería El Golpe, 2da. Sección, perteneciente al municipio de Cárdenas, en el estado de Tabasco. Esto debido a que cumple las características requeridas para la investigación, como son alta pobreza y marginación. El universo poblacional corresponde a los 116 hogares en los que viven los 500 habitantes de la localidad. El tamaño de muestra es de 90 hogares, y se determinó mediante muestreo aleatorio simple, considerándose un intervalo de confianza del 95%, y un margen de error del 5%.

La técnica para recolección de la información fue diseñada a partir de los objetivos de la investigación, habiéndose aplicado a los sujetos de estudio un cuestionario con preguntas cerradas, construido en base a investigaciones precedentes con características similares, así como en el cuestionario utilizado en el censo de ingreso-gasto aplicado por el INEGI en 2018. De igual manera, se tomó como referencia el manual de Metodología para la elaboración de estrategias de desarrollo local de Silva y Sandoval, publicado por la CEPAL en el año 2012. Para el análisis cuantitativo de la investigación se aplicó estadística descriptiva.

3. RESULTADOS

Las características sociodemográficas de las comunidades en situación de pobreza en México se corresponden con la transición demográfica presente en los países latinoamericanos: presencia mayoritaria de población joven y adulta, bajo nivel de escolaridad, precariedad del empleo y por lo tanto bajo nivel de vida de sus habitantes. Lo anterior se corresponde con los resultados obtenidos en la comunidad sujeto de estudio, donde el 20% son niños, 27% adolescentes y jóvenes, 42% son adultos y 10% adulto mayor Como puede observarse, está presente un bono demográfico sustancial susceptible de aprovecharse para el desarrollo local (figura 2).

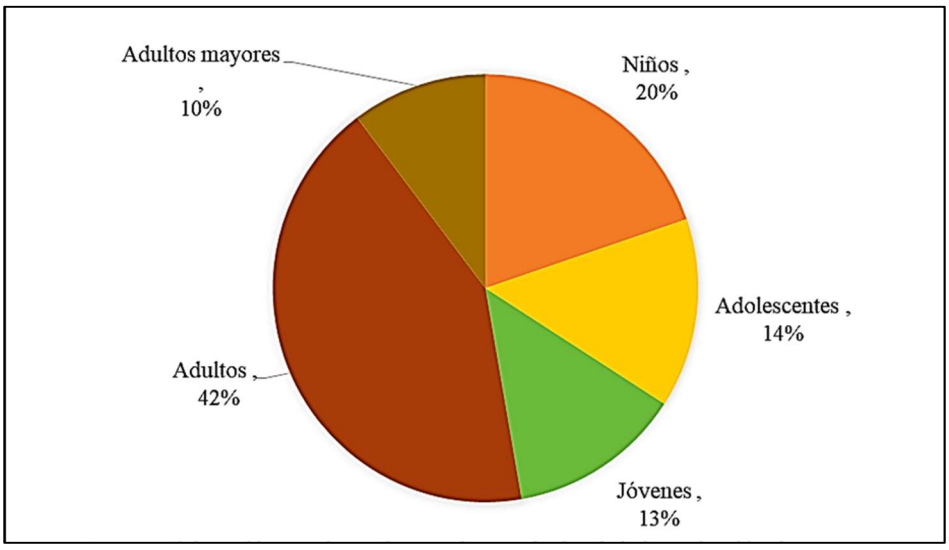


Figura 2**.** Distribución de la población
Fuente: Elaboración propia con base en los resultados de la investigación de campo.

Debido a las condiciones de marginación y su bajo nivel de escolaridad, el trabajo desarrollado por la población es precario y de bajo nivel agregado. De esta manera, las ocupaciones de la población masculina entrevistada están centradas en las actividades propias del campo y en la generación de productos del sector primario, los cuales se comercializan en la comunidad y en pocos casos tienen alcance externo a la misma. También se registran actividades de ganadería y pesca a baja escala. El sector servicios comienza a cobrar importancia a través del comercio y los servicios técnicos de carpintería, plomería, electricistas, mecánicos y veterinarios. En el caso de las mujeres, éstas se dedican generalmente a labores del hogar, actividad que en la mayoría de los casos no implica remuneración alguna. Asimismo, existe un alto nivel de desocupación (Figura 3).

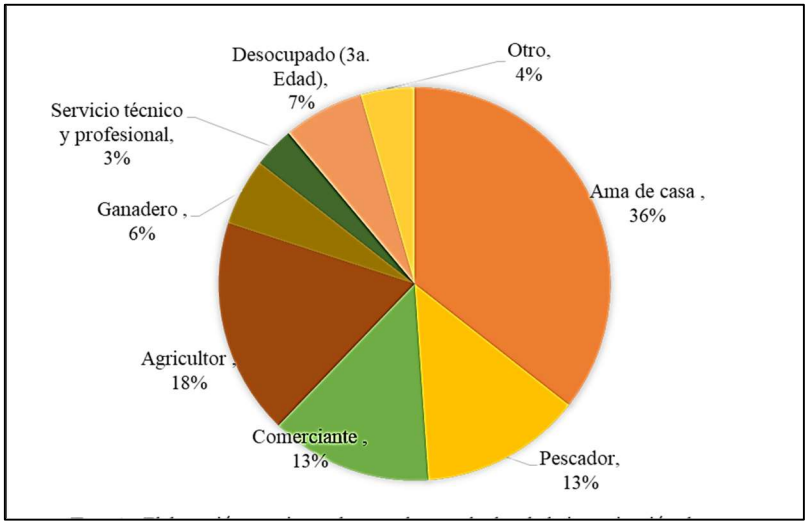


Figura 3. Ocupación y empleo
Fuente: Elaboración propia con base en los resultados de la investigación de campo.

Dada la precariedad de sus actividades económicas y su baja escolaridad, es natural que los ingresos de la población sean insuficientes para satisfacer sus necesidades básicas, no solo alimenticias sino también de educación, salud y vivienda. Así, se encontró que el 52% de los entrevistados percibe menos de 500 pesos semanales por su trabajo principal, el cual distribuido en un hogar con un promedio de 5 personas obliga a una vida precaria y limitada, pues dicho ingreso se usa principalmente para gastos de alimentación (65%), educación (20%), salud y medicamentos (8%) y vestimenta (5%). Asimismo, en el 60% de las familias encuestadas la figura paterna aporta el ingreso total del hogar, mientras que el 40% restante es aportado por diferentes miembros de la familia. Por las actividades extras que realizan y los apoyos gubernamentales que perciben las familias, el ingreso familiar semanal se incrementa (Figura 4).

Como puede apreciarse, los apoyos gubernamentales de transferencias monetarias han incrementado notablemente el ingreso familiar de los habitantes en situación de pobreza, lo que ha coadyuvado a reducir las carencias alimentarias. El impacto de los programas de apoyo que están beneficiando a los habitantes de la comunidad estudiada es relevante. Así, el de sembrando vida apoya al 32% de las familias; becas de educación básica un 40%; becas de educación media superior 19%; adultos mayores 13%; jóvenes construyendo el futuro 10%; becas de educación superior 6%, y personas con discapacidad 3%.

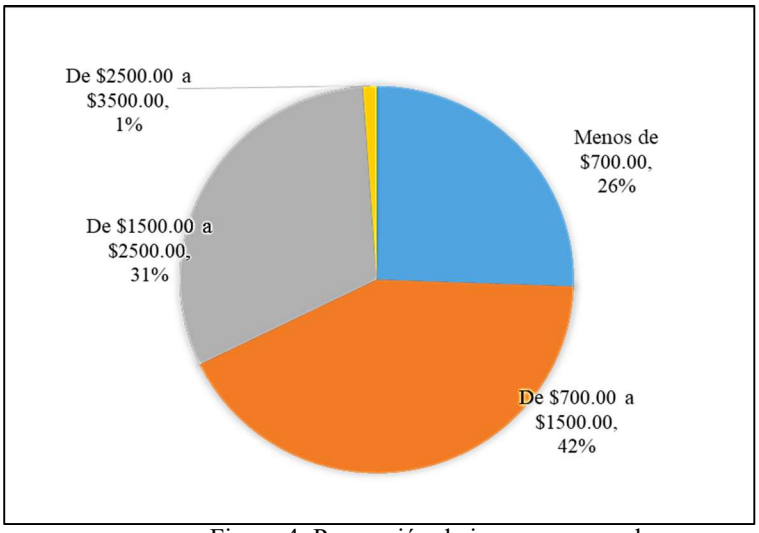


Figura 4. Percepción de ingreso semanal
Fuente: Elaboración propia con base en los resultados de la investigación de campo.

Un signo distintivo de la pobreza y marginación de la población es la falta de acceso a diversos bienes públicos básicos. Las carencias sociales se refieren a la privación del acceso a capacidades básicas para la vida y el bienestar como son los servicios básicos de educación, salud y vivienda. Así, en la comunidad estudiada la mitad de la población apenas si cuenta con estudios de nivel primaria, el 30% de secundaria, el 12% de preparatoria, y solo el 2% estudió alguna profesión. Así mismo, el 6% es analfabeta, pues no sabe leer ni escribir (Figura 5).

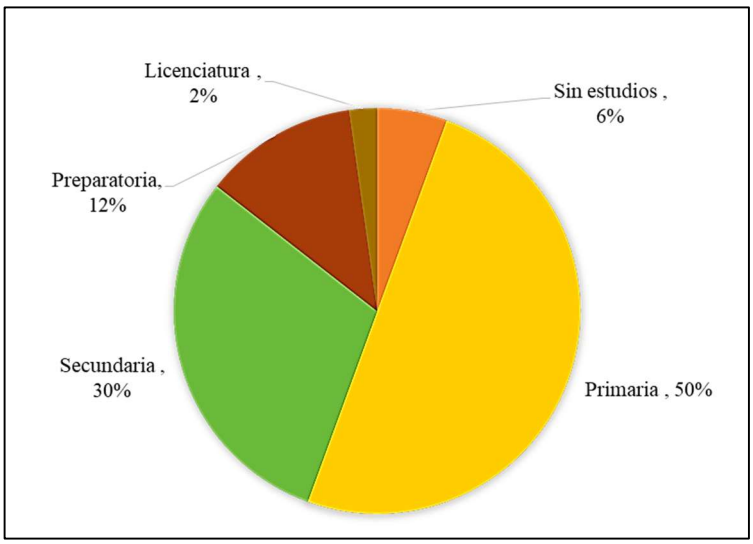


Figura 5. Escolaridad
Fuente: Elaboración propia con base en los resultados de la investigación de campo.

Respecto al acceso a servicios de salud, el 97% de la población está afiliada al programa denominado Seguro Popular de Salud (SPS), mismo que garantiza un paquete de servicios básicos y no supone seguridad social, y

únicamente el 3% accede a las prestaciones del Instituto Mexicano del Seguro Social (IMSS). Quienes reciben atención médica del SPS en el centro de salud de la comunidad vecina opinaron que el servicio es deficiente y limitado en la prestación del servicio y el surtimiento de medicamentos recetados, mientras que quienes reciben atención en el IMSS manifestaron estar satisfechos con el servicio.

En materia de vivienda, el 96% de los habitantes de la comunidad cuentan con casa propia, aunque las características de las mismas son precarias, pues la mayor parte de ellas apenas si cuentan con dos habitaciones, y algunas son de techo de lámina y piso de tierra, lo que hace a las familias que las habitan más vulnerables a las inclemencias del tiempo y a enfermedades (figura 6).

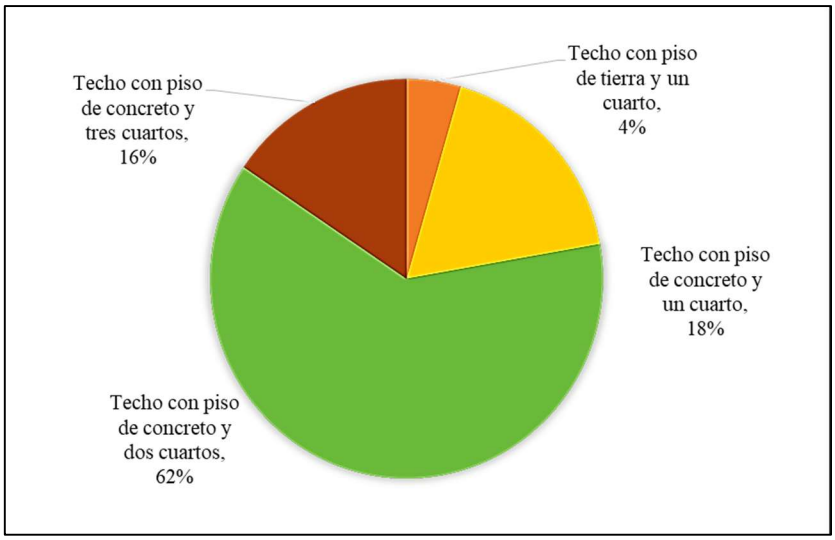


Figura 6. Características de la vivienda
Fuente: Elaboración propia con base en los resultados de la investigación de campo.

Respecto a la disponibilidad y acceso otros servicios básicos para el bienestar y el desarrollo local, el 90% de la población dispone de los servicios de energía eléctrica y agua potable, aunque su suministro es deficiente e irregular. Ambos servicios están relacionados: si no hay energía eléctrica, el pozo de distribución de agua potable no funciona. Al momento que se aplicó la encuesta, la comunidad tenía 20 días sin agua potable debido a fallas en la bomba de distribución. Por otra parte, el 91% de la población entrevistada coincidió en que los caminos y carreteras que comunican a la localidad se encuentran en mal estado y cada vez se deterioran más, sin que ninguna autoridad atienda este problema. Por lo mismo, el transporte público es deficiente e insuficiente para satisfacer las demandas de traslado de los habitantes de la comunidad.

Un aspecto relevante para el desarrollo local es la percepción que tiene la sociedad acerca de su propia realidad y la manera de superar sus problemas. Con tal propósito, el trabajo de campo incluyó un apartado especial. Al respecto, todos los entrevistados consideraron que la comunidad cuenta con importantes recurso naturales susceptibles de aprovechamiento, como son amplias extensiones de tierras fértiles para la agricultura y abundantes manglares y lagunas para la pesca y ostricultura, así como la elaboración de artesanías y la

carpintería. Otro aspecto potencial de desarrollo sería el emprendimiento del turismo ecológico sustentado en la riqueza natural que rodea a la comunidad. Al respecto, debe destacarse la satisfacción de la población con los programas públicos implementados por la administración federal actual. Así, la tierra fértil está siendo aprovechada para la reforestación a través del programa sembrado vida, del cual el 32% de la población está resultando beneficiada, ya sea porque reciben un ingreso en su calidad de propietarios, o porque cuentan con un empleo y reciben un ingreso como trabajadores agrícolas campesinos (Figura 7).

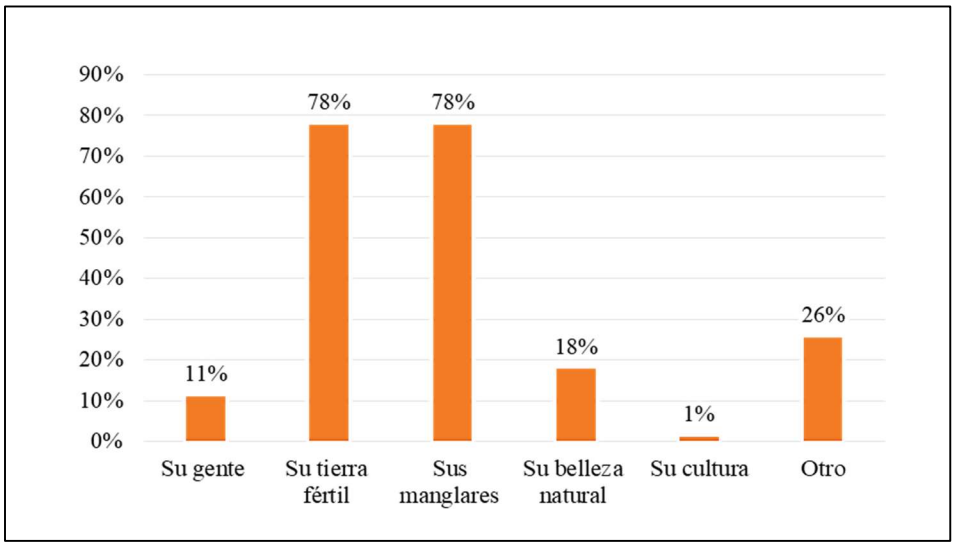


Figura 7. Percepción sobre los recursos aprovechables
Fuente: Elaboración propia con base en los resultados de la investigación de campo.

EL NUEVO ESCENARIO POR COVID-19

La pandemia de coronavirus SARS-CoV-2, que provoca la enfermedad COVID-19 continúa presente a escala mundial, y para mediados de septiembre de 2020 ya había afectado a más de 29.9 millones de personas en todo el mundo, en tanto que la cantidad global de fallecimientos estaba por arriba de los 942.000 habitantes, habiéndose Recuperado de: la enfermedad más de 19.9 millones de personas. Estados Unidos es el país más afectado, con más de 6.6 millones de contagios y más de 196,000 decesos por esta causa, seguido de India, que ya supera los 5.1 millones de casos, y 83,000 defunciones. Por su parte, Brasil rebasa ya los 4.4 millones de casos, así como más de 134,000 decesos; Rusia ha superado ya el millón de contagios, y Perú y Colombia sobrepasan los 700,000 infectados. Asimismo, México contaba ya con más de 670,000 casos, y España por su parte, acumulaba más de 614,000 casos y más de 30,000 muertes. Debido a la velocidad de expansión de la pandemia y su alta tasa de contagio, más de la mitad de la población mundial ha sido sometida a algún tipo de medida de política pública sanitaria para enfrentarla de manera más eficiente. Entre estas medidas destacan el confinamiento y distanciamiento social, con lo que los desplazamientos por todas las vías han quedado paralizados al igual que la actividad económica, lo que ha provocado una profunda recesión económica en todo el mundo (CONEVAL, 2020).

Más recientemente, al 03 de noviembre de 2021 a nivel mundial ya se reportaban 247 millones 472 mil 724 casos confirmados y 5 millones 012 mil 337 defunciones por esta causa. Por su parte, en México se habían confirmado 3 millones 814 mil 453 casos y 288 mil 887 defunciones totales por COVID-19 (SSA, 2021). Debido a la velocidad de expansión de la pandemia y su alta tasa de contagio, más de la mitad de la población mundial ha sido sometida a algún tipo de medida de política pública sanitaria para enfrentarla de manera más eficiente. Entre estas medidas destacan el confinamiento y distanciamiento social, con lo que los desplazamientos por todas las vías han quedado paralizados al igual que la actividad económica, lo que ha provocado una profunda recesión económica y por lo tanto mayor desempleo, pobreza y desigualdad en todo el mundo (CONEVAL, 2020).

En México, la emergencia sanitaria causada por el COVID-19 nos conduce a todos, sociedad en general, empresa, instituciones públicas y privadas y gobierno, a replantear las prioridades individuales y colectivas. El precio en vidas humanas y pérdidas económicas y sociales pagado hasta ahora, ha sido de una magnitud tal que no tiene comparación con ninguna crisis anterior. Ante ello, es urgente la necesidad de promover la existencia efectiva de un sistema de protección social en salud y seguridad social que garantice el bienestar de la población, sobre todo la más vulnerable y marginada, garantizando así el acceso a los derechos sociales en México. Como se ha enfatizado anteriormente, la pandemia se presenta en un contexto complicado de pobreza y vulnerabilidad, tanto en lo que corresponde a infraestructura sanitaria como a las condiciones de salud de la población mexicana, caracterizadas por la presencia de enfermedades crónico degenerativas, principalmente diabetes e hipertensión, así como sobrepeso y obesidad, todo lo cual magnifica el impacto de la enfermedad Covid 19 en términos de vidas humanas.

Al igual que en las circunstancias actuales, otras crisis económicas que han tenido lugar en México, han exhibido que no obstante ser un país de ingreso medio, los inaceptables niveles de pobreza y rezago socioeconómico presentes, hacen más evidente la vulnerabilidad social y económica. Una realidad presente en las familias mexicanas es su vulnerabilidad económica, toda vez que en su gran mayoría dependen directamente de sus ingresos laborales. De acuerdo a datos del Coneval (2020), el ingreso por trabajo subordinado corresponde a más de 50 por ciento del ingreso corriente total de los hogares, y cuatro de cada diez personas en México se encuentran en situación de pobreza, es decir, el 41.9 por ciento de la población total. Si bien, entre 2008 y 2018 en México se logró disminuir la situación de pobreza en 2.5 puntos porcentuales, pasando de 44.4 a 41.9 porciento, se estima que la pandemia tendrá un impacto drástico en materia de pobreza y desigualdad económica y social (Coneval, 2020).

En síntesis, México está enfrentando esta crisis en condiciones de vulnerabilidad sanitaria, económica y social. Entre otros determinantes de esta compleja situación se encuentran la alta prevalencia entre la población de diabetes y enfermedades cardiovasculares, la precariedad laboral, problemas de acceso al agua, pobreza y hacinamiento así como múltiples brechas de acceso a derechos sociales: salud, alimentación, educación,

vivienda y seguridad social, todo lo cual dificulta la adopción generalizada de las medidas preventivas recomendadas (Coneval, 2020). No obstante, la complejidad del contexto en el cual se manifiesta la pandemia en México, desde su inicio ha prevalecido una estrategia de acción encabezada por el gobierno federal a través de la Secretaría de Salud, misma que ha permitido disminuir el ritmo y cantidad de contagios, así como minimizar la pérdida de vidas humanas. Cabe mencionar que dicha estrategia se ha soportado enteramente en el conocimiento científico prevaleciente. Entre las medidas sanitarias adoptadas se encuentran el confinamiento social; la reconversión hospitalaria; la aplicación de pruebas para detectar personas contagiadas, y la contratación y capacitación de personal médico y paramédico para enfrentar la complejidad de la enfermedad.

4. CONCLUSIONES

La situación de pobreza y rezago social en México son problemas estructurales complejos. En el año 2018, el 41.9% de la población total se encontraba en pobreza multidimensional, y lo que es más grave, el 7.4% en pobreza extrema, sobre todo en las regiones rurales. Asimismo, los indicadores señalan que Tabasco se encuentra por encima de la media nacional en pobreza extrema, que lo colocan como uno de los estados más vulnerables en materia de pobreza y rezago social a nivel nacional. De ahí el interés por impulsar el desarrollo local a efectos de mejorar las condiciones de vida de los habitantes de las áreas rurales, a partir del aprovechamiento estratégico de las potencialidades que estos territorios poseen.

En la comunidad estudiada se encontró que las familias rurales perciben ingresos insuficientes para satisfacer sus necesidades básicas, motivo por el cual permanecen en situación de pobreza extrema, y son a la vez las que más privaciones sociales presentan, lo que les limita su derecho a una vida digna. Manifiestan también un importante rezago en materia de educación, salud y seguridad social, así como múltiples carencias en la calidad de su vivienda, limitaciones en el acceso a la energía eléctrica, servicios de agua potable, carreteras y transporte público.

Por otro lado, los habitantes de la comunidad cuentan con una perspectiva de desarrollo local e identifican claramente los recursos potenciales disponibles y su posible aprovechamiento, lo que facilita la implementación de un proceso de planeación estratégica. Así, identificada la percepción de la comunidad respecto a su circunstancia actual, así como las principales debilidades, fortalezas y potencialidades para el desarrollo local, resulta pertinente la elaboración de un plan estratégico de acción que debidamente aplicado posibilite en el mediano y largo plazo el desarrollo local, contando con la participación activa de la propia población y sus autoridades, el Estado en todos sus niveles y las organizaciones públicas y privadas que deban concurrir.

La planeación estratégica del desarrollo local significa una alternativa viable para la superación de la pobreza y el rezago social y económico de las comunidades marginadas del desarrollo nacional y regional. Por ello, la investigación concluye en la elaboración de un plan estratégico para impulsar el desarrollo local de las comunidades en situación de pobreza en Tabasco, mismo que a la vez es una propuesta metodológica para el desarrollo local aplicable en el ámbito nacional e internacional.

El plan estratégico de desarrollo local diseñado a partir de la investigación pretende impulsar el desarrollo local en la Ranchería el Golpe 2ª Sección, así como coadyuvar a reducir las condiciones de pobreza de sus habitantes. El plan está estructurado con los siguientes elementos: 1) Misión; 2) Visión; 3) Valores; 4) Objetivo general; 5) Análisis FODA; 6) Objetivos estratégicos, estrategias y líneas de acción; 7) Metas; y 8) Indicadores de desempeño.

Se puede concluir que el desarrollo y la superación de la pobreza es un proceso dinámico de mediano y largo plazo. Por lo tanto, la propuesta no constituye un plan terminado e inamovible, sino que el plan estratégico propuesto puede y debe irse actualizando y mejorando durante su implementación y con el paso del tiempo. Debe además ser adaptado a las especificidades propias de las regiones estudiadas. A partir del plan estratégico presentado se diseñarán los programas y proyectos necesarios para hacer realidad la visión de desarrollo local propuesta. Los objetivos estratégicos se derivan de la complejidad de la problemática que se desea resolver, como lo es la situación de pobreza en la comunidad de estudio.

REFERENCIAS

ABREU, J. El método de la investigación. **International Journal of Good Conscience**, 195-204. 2014.

ALCAÑIZ, M. El desarrollo local en el contexto de la globalización. **Revista de ciencias sociales Convergencia**, vol. 15, núm. 47, mayo-agosto, 2008, pp. 285-315. 2008.

ARMIJO, M. ***Manual de planificación estratégica e indicadores de desempeño en el sector público,*** ILPES/CEPAL, Chile. 2011.

BANCO MUNDIAL. Pobreza. Panorma Mundial. Disponible en https://www.bancomundial.org/es/topic/poverty/overview. 2019.

BOISIER, S. **Desarrollo (local)**: ***¿De qué estamos hablando? Transformaciones globales, Institucionales y Políticas de desarrollo local.*** Homo Sapiens, Rosario. 2001.

BRUE, S., & GRANT, R. ***Historia del pensamiento económico***. CENGAGE Learnning, México. 2009.

CHANDLER, A. ***Strategy and Structure. Chapters in the history of the American Industrial Enterprise***. Beard Books, New York. 2003.

CHAPOY, D. B. ***Planeación, programación y presupuestación***. UNAM, México. 2003.

CONEVAL.¿Qué es la medición de la pobreza? Disponible en https://www.coneval.org.mx/Medicion/MP/Paginas/Que-es-la-medicion-multidimensional-de-la-pobreza.aspx. 2018.

ESPINOZA, J. Estado social (de Derecho) en México. Una óptica desde el garantismo jurídico-social. **Revista Iberoamericana de Derecho Procesal Constitucional**, 9, 61-83. 2008.

GARCÍA, J. **Aproximación al estudio de la pobreza en México: Propuesta de política de estado contra la pobreza.** Universidad Juárez Autónoma de Tabasco, Tabasco, México. 2016.

HERNÁNDEZ, R., FERNÁNDEZ, C., & BAPTISTA, P. ***Metodología de la investigación***. MacGrawHill, México. 2014.

HERRERA, F. Enfoques y políticas de desarrollo rural en México: Una revisión de su construcción institucional. **Gestión y política pública**. Disponible en http://www.scielo.org.mx/scielo.php?script=sci_arttext&pid=S1405-10792013000100004#nota. 2013.

HUESO, A. y CASCANT, M. J. Metodología y técnicas cuantitativas de investigación. **Cuadernos docentes en procesos de desarrollo** (1) 81. 2012.

KRUGMAN, P., & WELLS, R. ***Macroeconomía.*** Reverté, México. 2014.

MIGUEL, A. ***Ciencia regional: Principios de economía y desarrollo***. EUMED, España. 2004.

ONU. *Desarrollo.* Disponible en: https://www.un.org/es/sections/what-we-do/promote-sustainable-development/ 2012.

ONU. La Asamblea General adopta la Agenda 2030. Objetivos de Desarrollo Sostenible. Disponible en https://www.un.org/sustainabledevelopment/es/2015/09/la-asamblea-general-adopta-la-agenda-2030-para-el-desarrollo-sostenible/. 25 de septiembre de 2015.

PIKETTY, T. ***El capital en el siglo XXI*.** Fondo de Cultura Económica, México. 2014.

PORTALES, L. Los pobres como agentes de su desarrollo, la lucha contra la pobreza y la exclusión desde lo local. **Revue Interventions Économiques**, 51, 1–18. Disponible en https://journals.openedition.org/interventionseconomiques/2172. 2014.

RAMÍREZ, A., SÁNCHEZ, J., & GARCÍA, A. El desarrollo sustentable: interpretación y análisis**. Revista del Centro de Investigación. Universidad La Salle,** vol. 6, núm. 21, julio-diciembre, 2004, 55-59. Disponible en https://www.redalyc.org/pdf/342/34202107.pdf. 2004.

ROLDÁN, N. Operan más de 5 mil programas sociales con gasto millonario, pero no logran disminuir la pobreza. **Animal Político.** Disponible en https://www.animalpolitico.com/2017/04/programas-sociales-pobreza. 2017.

SAMUELSON, P., & NORDHAUS, W. ***Economía con aplicaciones para Latinoamérica***. McGraw-Hill, México. 2010.

SILVA, I., & SANDOVAL, C. ***Metodología para la elaboración de estrategias de desarrollo local.*** CEPAL, Chile. 2012.

SURDAY, Y. El análisis de información y las investigaciones cuantitativa y cualitativa. **Revista Cubana de Salud Pública**, 33 (2) 11. 2007.

ZARTA, P. La sustentabilidad o sostenibilidad: un concepto poderoso para la humanidad.**Tábula Rasa,** 28, 409-423. Disponible en http://www.scielo.org.co/pdf/tara/n28/1794-2489-tara-28-00409.pdf. 2018.

Capítulo 3 pp 31-46

LA SELECCIÓN DE MUESTRA DE POBLACIONES CON ESTRUCTURA DE RED: MODELOS ALTERNATIVOS PARA EVALUAR ASPECTOS DE UNA PANDEMIA

Carlos Bouza[1], Sira Allende[1], Faizan Danish[2] and S.E.H. Rizvi[3]

[1]Universidad de La Habana, Cuba
[2]Research Consultation Services, Doha, Qatar
[3]Division of Statistics and Computer Science, Faculty of Basic Sciences, SKUAST-JAMMU, (J&K), India

Resumen
En este trabajo presentamos modelos de muestreo que permiten hacer estudios en redes. Estos modelos permiten sostener investigaciones durante el desarrollo de pandemias y caracterizar las dinámicas. Se discute como evaluar el posible efecto de políticas sanitarias en pandemias como la del COVID19.

1. INTRODUCCIÓN

En varios modelos epidemiológicos se asume una compartimentación y que esta permite considerar que los individuos se mezclan en forma homogénea dentro de cada grupo (clúster, estrato). Al comenzar una epidemia un pequeño número de infestados aparecen y la transmisión se desarrolla a partir de unas probabilidades que fijan los patrones de los contactos entre los individuos de la población. Es importante poder establecer un modelo que nos dé una descripción de este patrón. Desde este punto de vista el problema requiere determinar una función de distribución de probabilidad. Vea Diekmann-Heesterbeek (2000). Un patrón observado en el caso del COVID19 es que algunos portadores transmiten la infección a muchos de sus contactos y que una cantidad significativa de ellos no la transmite con la misma frecuencia. Así que asumir que los portadores se mezclan con los susceptibles homogéneamente no parece producir un adecuado patrón. En la epidemia del SARS, del 2002–2003, se observó esto, Ferguson (2005), Longini (2005). Además, se observó en esta epidemia que el valor reproductivo fue mucho menor que lo previamente supuesto.

Cuando el patrón es heterogéneo el podemos modelar el proceso mediante una red en la que un gafo, que tiene los individuos como nodos y los contactos como aristas, representa el fenómeno.

La Teoría de Grafos es utilizada en diversos contextos, especialmente en la era de la informática, para describir redes de computadoras, sociales, de comunicación etc., vea Lloyd (2001), Ahnet al. (2007), Benevenuto et al. (2009), Kwak et al. (2010), Mislove et al. (2007), Sala et al. (2010), Wilson et al. (2009). En epidemiologia esta también es útil usando varios modelos: grafos unidireccionales, bidireccionales, no dirigidos, etc. Véase su uso en Meyers et al (2005), Newman (2002), Keeling - Eames (2006). Es lógico considerar que una arista representa el contacto entre individuos, si lo hubiere. La Teoría de Grafos permite evaluar diversos aspectos del comportamiento de la dinámica de la epidemia. Cuando esta comienza se transmite a través de contactos que van de los vértices (nodos, individuos) contaminados a otros. Al haber un brote se activan algunas aristas (contactos) y el tamaño de la infestación a es el clúster de los nodos conectados por una cadena continua de aristas.

El uso de herramientas de la Teoría de Grafos permite hacer predicciones de las probabilidades de contagio de un individuo, determinar su riesgo de estar en una cadena de contaminados etc.

El seguimiento de los recuperados atrae mucha atención. En la literatura reciente se reportan estudios retrospectivos de las características clínicas de pacientes infectados con SARS-CoV-2. Por ejemplo, el estudiar las variables definidas por los valores de Ct, anti-SARS-CoV-2 IgG, anticuerpos IgM en recuperados del COVID-19, la diferencia de anti-SARS-CoV-2 IgG con anticuerpos IgM. Vea An et al. (2020).
En la próxima sección se presentan algunos de los métodos generales para enfocar los estudios de grafos usando muestras. Estos se dividen considerando si el estudio se desarrolla sobre un grafo estático o sobre uno temporal. Le sigue una presentación de algunos de los más populares algoritmos de selección que permiten conformar muestras aleatorias utilizando la estructura de grafos que modela las relaciones en una población. Se brindan los pseudo códigos de estos y se dan referencia para la búsqueda de códigos en los repositorios de libre acceso. Finalmente, en la sección 3 se presentan métodos inferenciales para establecer propiedades de los grafos como los de determinar un grafo que represente el grafo total permitiendo establecer el comportamiento de políticas usándole. La determinación de distribuciones de probabilidad es discutida con detalle.

2. POBLACIONES CON UNA ESTRUCTURA DE GRAFOS

Los usuarios acuden al estadístico para tratar de obtener respuestas sobre el tamaño de la muestra, como hacer la selección de esta y como obtener adecuadas estimaciones de parámetros de la población de su interés. Las preguntas que hacen vienen generalmente en la forma siguiente:

- ¿Qué métodos de muestrear es bueno?
- ¿Cuáles son las bondades del método de muestreo?
- ¿Qué tamaño debe tener la muestra?
- Como mido la eficiencia de la muestra.

Estas preguntas son explícitas, pero al plantear el problema a estudiar podemos considerar que detecta que la población, con N individuos $U = \{u_1, \dots, u_N\}$, es tal que se pueden establecer conexiones entre ellos y que podemos representarla mediante un grafo. Sencillamente tomando

$$I(u_i, u_j) = \begin{cases} 1 & si\ u_i\ se\ relaciona\ con\ u_j \\ 0 & en\ otro\ caso \end{cases}$$

podemos determinar un grafo G: los individuos son representados por nodos y la existencia de relaciones entre ellos son las aristas.

En muchas aplicaciones el uso de esta estructura de grafos ayuda en la modelación de los fenómenos. Tal es caso en simulaciones de protocolos de navegación de internet, la propagación de virus, el comportamiento de políticas de inmunización etc. Por ello en estudios de transmisión del COVID puede ser recomendable usar métodos de muestreo que, tomando en cuenta esta estructura permita obtener información sobre las relaciones entre individuos al medir variables de la transmisión y estimar los parámetros que sean de interés como los asociados a riesgos, importancia de las fuentes de transmisión y otros factores de interés epidemiológico.

Veamos algunos patrones que esperamos satisfaga, un grafo estático y original G, que nos permitan estudiar tales poblaciones. Esencialmente estos se tratarían a partir de hacer un adecuado escalamiento. Los patrones se establecen al determinar una distribución. Esta será determinada por los intereses del epidemiólogo. Veamos algunos que serían de importancia en el estudio del comportamiento del COVID19.

• S1: Distribución de los grados de entrada (in-degree) para cada grado. Su estudio proporciona herramientas para estudiar tasas de contaminación. Para un individuo esto permite utilizar la información proporcionada sobre cuantos le contactaron. Tal es el caso al desear observar los contactos de portadores comprobados.

Modelamos esto al considerar que tenemos que evaluar la variable

$$s1(u_j \to u_i) = \begin{cases} 1 \; si \;\; u_j \; posee \; el \; virus \; y \; contacta \; a \; u_i \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad en \; otro \; caso \end{cases}$$

Su grado de entrada es

$$ge(u_i) = \sum_{j \in s} s1(u_j \to u_i)$$

Como se distribuyen los tipos de individuos $ge(u_i) = t$ es de interés para establecer la probabilidad de que un individuo de la población sea contactado por t contagiados: $P(u = t)$. Esto llevaría a buscar la distribución de la correspondiente variable aleatoria. Con ella, por ejemplo, se podrá estudiar el efecto de la tasa de contaminados con que tuvo contacto en términos de la proporción del número de contactos de los individuos en la población. Midiendo

$$p_{t\downarrow} = \frac{1}{\|s(t\downarrow)\|} \sum_{i \in s(t\downarrow)} ge(u_i);\; s(t\downarrow) = \{u_i \in s | ge(u_i) = t\}$$

podemos determinar la distribución correspondiente. Esto permite al epidemiólogo trabajar sobre las tasas de contagio y su efecto a partir de determinar una aproximación a la ley de probabilidad que les genera. Generalmente el modelo teórico de esta es una ley de la potencia o por una CDF con colas pesadas. Vea discusión en Gjoka et al. (2010).

• S2: Distribución de los grados de salida (out-degree) para cada grado. Su estudio proporciona herramientas para estudiar tasas de contaminación. Para un individuo esto permite utilizar la información proporcionada para establecer con cuantos contactó. Por ejemplo, en estudios retrospectivos de los contactos que pudieron ser la fuente de su contaminación con el virus con SARS-CoV-2. Así se podrá estudiar la tasa de contaminados con el virus generada por haberle tenido como contacto. Esto permite determinar otra distribución que permite obtener ideas sobre las tasas de contagio y su efecto. En este caso se computa

$$s2(u_j \leftarrow u_i) = \begin{cases} 1 \; si \;\; u_i \; posee \; el \; virus \; y \; contacta \; a \; u_j \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad en \; otro \; caso \end{cases}$$

También contamos el número de nodos con estos grados

$$gs(u_i) = \sum_{j \in s} s2(u_j \leftarrow u_i)$$

y computamos

$$p_{t\uparrow} = \frac{1}{\|s(t\uparrow)\|} \sum_{i \in s(t\uparrow)} ge(u_i);\; s(t\uparrow) = \{u_i \in s | gs(u_i) = t\}$$

También generalmente el modelo probabilístico es descrito por una ley de la potencia o por una CDF con colas pesadas. Vea discusión en Gjoka et al. (2010)

• S3: Distribución de los tamaños de los componentes conectados débilmente ("WCC"). Para individuos observamos las conexiones del tipo

$$s3(u_i \sim u_j) = \begin{cases} 1 \; si \; se \; relacionan \\ 0 \; si \; no \; lo \; hacen \end{cases}$$

El interés lleva a identificar los individuos que cumplen con esta relación, que determina una arista no dirigida. Tal es el caso cuando se hace un estudio de colectivos, se detecta la existencia de contacto entre ellos. Podemos usar esto para detectar lo que en Cuba se ha dado en llamar "araña" en los estudios de las cadenas. En muchos

casos estas no forman caminos largos sino conexiones entre un grupo de individuos muy cercanos. En este caso medimos para cada individuo en la muestra

$$a(u_i) = \sum_{j \in s} s3(u_i \sim u_j)$$

y computamos

$$p_{t(a)} = \frac{1}{\|sa_t\|} \sum_{i \in sa_t} a(u_i);\ sa_t = \{u_i \in s | a(u_i) = t\}$$

Así que se computaría el número de los pares de nodos entre los que existe un camino no dirigido con t aristas.

• S4: Distribución de los tamaños de los componentes conectados fuertemente ("SCC"): Cuando tenemos que los contactos de un individuo determinan una cadena el interés, como la de los contagios. Utilizamos

$$s1(u_j \to u_i) = \begin{cases} 1\ si\ \ u_j\ posee\ el\ virus\ y\ contacta\ a\ u_i \\ 0 \qquad\qquad\qquad\qquad\qquad\qquad en\ otro\ caso \end{cases}$$

el interés es determinar un par de nodos entre los que existe un camino dirigido que los conecta. Al epidemiólogo le interesará la longitud del camino entre dos individuos y u_h. La denotamos $L(u_i, u_h)$. En epidemias es de importancia conocer cuántos y quienes son los individuos "fuertemente conectados" en un camino de longitud t. Ejemplificando tomemos el caso cuando A contamina a B, B a C y C a D. Hay un camino de A a D. En tal caso $s1(A \to B)$, $s1(B \to C)$ y $s1(C \to D)$ y L (A, D) = 3.
Al tener una muestra de tamaño n se puede determinar la proporción de componentes conectados fuertemente a través de caminos de longitud t al computar

$$p(C(t)) = \frac{n(t)}{2^n};\ \ n(t) = número\ de\ caminos\ de\ longitud\ t$$

Esta es la probabilidad de observar tal longitud entre todos los caminos posibles. Sin embargo, es mayor el interés en el número de estos entre los existentes. Si el número de estos es #C va a ser más útil al epidemiólogo establecer como estimación de la probabilidad $p(C(t)) = \frac{n(t)}{\#C}$.
Esta va a ser de importancia para estimar cuantos de estos caminos (cadenas d transmisión) que se esperarían si se seleccionan M individuos. Además, permitirá establecer las necedades de extender las pruebas entre asintomáticos del COVID19

• S5: Número de ploteos de los brincos (Hop-plots, HP). Este es el número de pares de nodos entre los cuales hay una distancia $t \leq h$, denotado por $P(h)$, h es el número de brincos, Benevenuto et al. (2009). Esto es de interés para estudiar como una persona es contaminada a partir haber t contactos previos en la cadena en la que está incluido. El epidemiólogo estaría interesado en muchas ocasiones en determinar con que frecuencia se espera tener nuevos casos a partir de un contagio incluido en una larga cadena de individuos.

• S6: Máximo de los HP (MHP). Este es el mayor de los componentes conectados débilmente (WCC). Es obvio su importancia y como la distribución obtenida en S1 permitirá estimar las esperanzas y las variabilidades de las variables epidemiológicas y de calidad de vida tras la recuperación.

• S7: La distribución de los coeficientes del clustering C_d . Este es el promedio de la fracción de las aristas que existen, ver Mislove, et al. (2007). Su cálculo se basa en que, para un nodo "v" con k vecinos hay a lo más $k(k-1)/2$ aristas. C_v denota cual es la fracción de las aristas que realmente existen

$$C_d = \frac{1}{\|C_{v(d)}\|} \sum_{v \in C_{v(d)}} C_v, C_{v(d)} = \{v | tiene\ grado\ d\}$$

Esta sirve para valorar las tasas de contactos. Además, permitirá establecer estratos una vez seleccionada la muestra. Sobre ellos, una vez identificados, los epidemiólogos podrán proyectar sus investigaciones de campo sobre las secuelas entre los recuperados del virus del COVID19, ante los grupos caracterizados por este clustering.

Para los grafos temporales sus patrones se miden en términos de la secuencia de los grafos en el tiempo. Estos van a ser necesariamente útiles cuando se estudie el COVID pensando en los rebrotes. En ellos determinaremos un valor para el grafo y analizamos cómo se comporta este en el tiempo. Los criterios se tratan a partir de las distribuciones de probabilidad estimadas. Lo deseable es partir de un tamaño de la muestra fijo n_0.

• T1: Ley de potencia de la Densificación (DPL), Gjoka (2010). El número de aristas versus el de nodos través del tiempo. DPL establece que $e(t) \propto n(t)_a$. El exponente de la densificación típicamente es mayor que 1, por tanto, el grado promedio de un nodo en el grafo crece con el tiempo. En estudios sobre el comportamiento de las redes formadas por los contagios, esto permitirá estudiar cómo se comportan las relaciones entre individuos (contagiados, en recuperación etc.) y las conexiones. En ellos se establecerían aspectos como la densidad de los contactos. Esto, por ejemplo, es importante cuando evaluemos cómo se comportan los contactos tras el confinamiento en el tiempo. Por ejemplo, se espera que en la primera etapa $(t = 1)$ esta no sea demasiado alta, al revaluar esto tras un periodo de tiempo $(t = 2)$ podríamos evaluar si esta crece, decrece y que en que mediada.

• T2: El efectivo diámetro del grafo a través del tiempo. Este es el número mínimo de brincos en el que él puede llegarse al 90% de los pares conectados. Este generalmente se reduce o estabiliza con el crecimiento del grafo en el tiempo, vea Gjoka, M. (2010). Note que en estudios sobre la magnitud de los contaminados asintomáticos esto es muy importante, pues nos permitirá establecer cuan amplia es la red de posibles contagiados y evaluar el comportamiento en distintos momentos.

• T3: El tamaño normalizado del mayor componente conectado $t(CC)$ a través del tiempo. Con esta medida se puede establecer cuan larga puede ser la máxima cadena de contactos con un individuo con el virus y evaluar cómo se comporta esta, en distintos momentos. Con ello el epidemiólogo va a poder discernir sobre diversos aspectos de la prevalencia y la difusión del virus.

• T4: promedio del coeficiente de clustering C a través del tiempo, Mislove, et al. (2007), $C = \frac{1}{\|G_t\|}\sum_{v\in G_t} C_{v|t}$.

3. ALGUNOS ALGORITMOS PARA MUESTREAR G

3.1. Algunas ideas.

Al enfrentarse un problema de muestreo en grafos se puedan usar diferentes diseños para generar un grafo muestral. Esto se basan en usar la estructura de grafo que caracteriza las conexiones de los individuos en una población. Entre ellas están el uso de criterio clustering y de particionamiento.

El problema que se encara al muestrear un grafo dirigido grande G, es crear un grafo pequeño G_S que sea similar a este, en términos de tener similares propiedades. Esto es, sabiendo que la población tiene estructura de grafo, como es el caso del COVID19, no sobeos como se conforman los caminos, las diadas, triadas etc. Tomando una muestra se observa G_S y se hacen generalizaciones sobre como se estructura la población y con ello el epidemiólogo podrá concebir campañas de control, vacunación, detección etc.

El epidemiólogo que realiza el estudio debe determinar si su interés es establecer cuan similar es el grafo muestral G_S al compararle con el original grafo G. Esto puede tomar forma al plantear que G_S debe ser similar a un grafo previamente observado G si este tuviese el mismo tamaño. Por ejemplo, con la observación de los resultados de la pandemia se puede considerar este como un grafo poblacional: el de los individuos que fueron detectados como infectados, posibles, asintomáticos etc. y sus contactos. En la literatura tenemos diversos métodos de búsqueda para establecer las relaciones entre G y G_S. Veamos algunos utilizables en el contexto de los estudios de la COVID19

- Meta del Escalamiento hacia abajo (Scale-down goal, SBG). Esta trata de que haya una buena coincidencia entre G y G_S.

Para generar G_S debemos tener un grafo estático y dirigido G con n nodos. Tomando la población determinada por la pandemia tendremos este grafo. El objetivo es crear un grafo muestral con n_0 nodos, $n_0 \ll n$, que sea similar a G. Este será más pequeño y por tanto más sencillo y barato será estudiarle. Esto es de especial importancia en el caso post-pandemia en el cuándo se desee establecer la existencia de secuelas, inmunidades etc. Claramente el epidemiólogo deseara tener un grafo G_S es que este tenga similares propiedades a G.

- Meta hacia atrás en el tiempo (Back-in-time goal, BTG). Esta trata de que haya coincidencias entre G y G_S en la evolución temporal.

Este procedimiento se basa en el interés de movernos en el tiempo hacia el pasado y tratar de imitar la versión que tenemos del grafo estático ya observado G. Tomando G_{n_0} (obtenido al final de la pandemia) como G, en un cierto momento pasado en que tenía n_0 nodos, la meta es hallar un grafo muestral obtenido de sobre los n_0 nodos que sea lo más similar posible a G_{n_0} . La evaluación de diversos aspectos del COVID19 se haría sobre G_S sabiendo que el efecto de las decisiones seria similares al llevarle al grupo de individuos que fue usado para determinar G_{n_0}.

Veamos primeramente como podemos determinar un método de muestreo que haga coincidentes G_S y G en términos de propiedades inherentes a los grafos. En lo que sigue cuando hablemos de selección aleatoria consideramos se utiliza un mecanismo que seleccione con una probabilidad uniforme. Es decir que es equivalente a un diseño del tipo muestreo simple aleatorio.

Poseer una buena representación de G va a permitir usarle en simulaciones para evaluar políticas de manejo del grafo real. Esto es de particular importancia en estudios de epidemiologia, comunicación social etc.

Un enfoque general es condensar el grafo buscando una mejor visualización, vea Ribeiro-Towsley (2010), Sala et al. (2010). Esto permitirá aumentar la velocidad de los algoritmos en la búsqueda de respuestas, vea Wang et al. (2010). Estudios de internet sugieren que, si el grafo es no dirigido algunas de las propiedades de los grafos, como separabilidad y estabilidad, se preservan usando una fracción de muestreo de los nodos no mayor del 0,3, vea Watts-Strogatz (1998).

2.2. Algunos procedimientos para seleccionar muestras

Podemos dividir los métodos de selección de muestras en tres grandes grupos:

- Selección aleatoria de nodos.
- Selección aleatoria de aristas.
- Técnicas de exploración.

Los dos primeros consisten en determinar la población bajo estudio en términos de los nodos (individuos) o de las aristas (conexiones, contactos) y seleccionar aleatoriamente n de ellos. La exploración no es sino una simulación de paseaos aleatorios que remede cómo el virus del SARS2 se propaga, y así determinar una muestra de nodos (quizás de individuos contagiados).

2.3. Selección aleatoria de nodos

Existen tres populares enfoques.

1. Crear un grafo muestral al seleccionar aleatoriamente n nodos de los N de la población. La muestra es el grafo determinado por estos n nodos. Este es conocido como Nodos Aleatorios (Random Node, RN). Estos métodos de selección no retienen la distribución conocida como la Ley de la potencia de los grados, ver Kurant et al. (2010). En general el interés del investigador estará centrado en determinar un método de muestreo tal que nos permita obtener una distribución adecuada de los grados. O sea, del número de contactos.

2. Asignar probabilidades desiguales a los nodos (PDN). La probabilidad de observar un individuo será proporcional a un cierto peso. Esto se basará en criterios del epidemiólogo que asignará mayores pesos a los individuos que considere sean más "importantes". Por ejemplo, el camarero de un restaurant va a ser más importante que un empleado administrativo de un hotel pues tiene más contacto con otras personas. En algunas aplicaciones esto es bueno y en otras no pues habrá un sesgo hacia los nodos con mayores grados.

3. Un método clásico es Muestreo Primer-Respiro (Breadth-First Sampling, BFS). Este es un muestreo de nodos. Vea detales en Kurant et al. (2010), Lee et al. (2006), Wilson et al. (2009), et al., Mislove, et al. (2007), Ahn et al. (2007). BFS permite hallar los nodos más cercanos el inicial y es usado para determinar distancias. Esto aparece como especialmente útil cuando estamos detectando posibles contactos con un portador del virus. Veamos cómo funciona el algoritmo

Fijar t=0
Seleccionar aleatoriamente una semilla v
Guardar los nodos muestreados en NM
Guardar los nodos procesados en NP (los seleccionados por el muestreo y los que se incorporaron por ser vecinos)
Mover v una vez procesado ponerle en NM
Todos los vecinos del nodo v se insertan en NP si no fue ya procesada
Mientras $t < n, t = t + 1$
Ir al paso 2

BFS tiende a preferir los nodos con grandes grados, consúltese Gjoka et al. (2010), pues estos serían visitados con mayor frecuencia. También obtiene altos valores de CC debido a este sesgo. Como fijamos el epidemiólogo va a estar interesado en tener para su estudio individuos con más contactos (grados).

4. El paseo aleatorio de Metrópolis-Hasting (MHRW). Es un algoritmo del tipo Markov-Chain Monte Carlo (MCMC) que genera aleatoriamente nodos de acuerdo a la distribución del grado, ver Gjoka (2010). Normalmente esto es y fácil de obtener usando muestreo directo. MHRW se basa en diseñar la CDF al aceptar o rechazar las propuestas. La función propuesta cambia las probabilidades transición permitiendo converger hacia la verdadera distribución. MHRW fue diseñado originalmente para grafos no dirigidos. Su adaptación para el caso dirigido fue desarrollada en Wang et al. (2010). Note que MHRW considera válido todos los

duplicados. Esto permite que la distribución de los nodos converja a la distribución uniforme. La distribución obtenida al usar MHRW es muy similar al del grafo original, ver Gjoka et al. (2010), Wang et al. (2010). Este se desempeña muy bien en grafos bien conectados. Así que cuando el epidemiólogo tenga tal muestra podrá confiar en que las decisiones que aporte a partir del estudio del grafo obtenido y de la distribución que obtenga debe ser adecuada para la población.

Este se implementa mediante el pseudo-código siguiente:

Fijar t=0 y n.

Seleccionar aleatoriamente un nodo v con grado diferente de cero

La función propuesta es $Q(v) = k_v$ (grado de v).

Determinar los vecinos de v : $C(v)$

Seleccionar aleatoriamente un $w \in Q(v)$ generando una variable Z de acuerdo a una uniforme $U(0,1)$.

Si $z < k_v/k_w$ se acepta la propuesta y el muestreo pasar a usar w

En otro caso se mantiene en v.

Cambiar t=t+1

A menos que $t > n$ ir al paso 2.

La probabilidad de transición es

$$P_{v,w} = \begin{cases} min\left(\frac{1}{k_v}, \frac{1}{k_w}\right) \; si \; u \; y \; w \; son \; vecinos \\ 1 - \sum_{w \neq v} min\left(\frac{1}{k_v}, \frac{1}{k_w}\right) \; si \; v = w \\ 0 \qquad\qquad\qquad\qquad en \; otro \; caso \end{cases}$$

Este algoritmo permite cambiar la probabilidad de transición y hay una convergencia a la verdadera probabilidad. Ver para detalles Wang et al (2010). Note que si el grado de w $(Q(w))$ es pequeño este tendrá una probabilidad pequeña de ser un candidato, pero una vez aceptado tendrá una alta probabilidad de ser incluido en la muestra. Este método permite rectificar el sesgo de otros procedimientos hacia los nodos con mayor número de grados.

5. Muestreo Adaptativo de Nodos. Sea una variable Y y un umbral A. Si tomamos un nodo y evaluamos si $Y_{ij} \geq A$ al considerar los demás individuos de la población podemos determinar el conjunto $\{u_j | Y_{ij} \geq A(1)\} = C(i1)$. Esto determina la conexión entre dos individuos en una estructura de grafos. Así que Y_{ij} sería el valor de la arista que los conecta. Si $u_j \in C(i1)$ se repite el proceso con sus vecinos y se determina $\{u_t | u_j \in C(i1), Y_{jt} \geq A(j)\} = C(i2)$. Se repite el proceso hasta que no se pueda generarse nuevos conjuntos C(ih). El conjunto $C(i) = \bigcup_k C(ik)$ es denominado cluster adatativo, ver Bouza (2000) y Thompson-Seber (1990). Como se ve este es un sub-grafo de la población que determina el grafo.

Si tomamos n individuos (nodos) tenemos un grafo dado por las conexiones presentes en los clústeres adaptativos y entre ellos. Los individuos $u_q \in C(ik)$ que pertenecen solo a ese clúster son denominados unidades fronteras.

Entonces al ser observado un individuo u_i esto puede ser debido a que

1. Se observa una de las $R(i)$ redes a las que pertenece.
2. Se observa al menos uno del $K(i)$ clústeres donde e s unidad frontera.

Así que la probabilidad de inclusión de ese individuo es

$$Prob\ de\ observar\ u_i = \pi_i = \frac{R(i) + K(i)}{N}$$

Si los contactos son aleatorios una variable Bernoulli modela la existencia de conexiones. Sea

$$B_{ij} = \begin{cases} 1 & si\ j \in C(i) \\ 0\ en\ otro\ caso \end{cases},\ \ E(B_{ij}) = Q_{ij}$$

Entonces

$$E(R(i)) = \sum_{j \in U \setminus u_i} Q_{ij} = Q(i)$$

En epidemiologia es de particular importancia la existencia de contactos entre portadores y susceptibles por lo que Y_{ij} sería una variable Bernoulli, $A = 1$. Si $A = 1$ hubo un contacto y el interés es saber cual es la esperanza de su suma (número de personas posiblemente contagiadas)

$$t(i) = \sum_{j \neq i} Y_{ij}$$

$$E(t(i)) = nQ(i)$$

la que es estimada por

$$\widehat{E(t(i))} = nR(i)/N$$

Para la población parece adecuado usar como estimador del número de posibles contagiados a

$$t = \frac{N}{n} \sum_{i \in s} \frac{1}{R(i)} \sum_{j \neq i} Y_{ij}$$

Bouza (2000) sugirió el uso de Bootstrap para hacer la estimación de la satisfacción de las hipótesis planteadas por Liu-Singh (1990) que

$$S(t) = \frac{1}{n-1} \left(\frac{1}{R(i)} \sum_{j \neq i} Y_{ij} - \frac{t}{N} \right)^2$$

Este es desarrollable en Series de Edgeworth.

2.4. Selección aleatoria de aristas

Los modelos basados en seleccionar aristas han probado su ineficacia amén de que en muchas ocasiones la conexión entre nodos no es conocida de antemano. En epidemiologia ese es el caso generalmente. La selección aleatoria de nodos se ha comprobado en la práctica es mucho mejor. Trabajan aún mejores métodos en los que se aprovechan la estructura subyacente en los grafos como los paseos aleatorios y los fuegos de bosque. Esos últimos se desempeñan muy bien con muestras donde la fracción de muestreo es menor del 0,15, ver Wang et al, (2010) y Wilson et al. (2009).
Esto puede ser descrito en forma similar a lo que se expuso para los nodos.

1. Dado un grafo con N aristas se seleccionan aleatoriamente n. Este es conocido como muestreo de Aristas Aleatorias (Random Edge, RE). Los grafos muestrales van a estar muy poco conectados, por lo que su diámetro será muy grande y no respetará la estructura de la comunidad.

2. Muestreo de Nodos-Aristas (Random Node- Edge, RNE). En este se selecciona aleatoriamente un nodo y entonces seleccionamos aleatoriamente una arista incidente a este. RE se puede considerar como sesgado para la estimación de los grados de los nodos, pues estos tendrán más aristas incidentes a él. RNE no lo tiene.

3. Una variante es la recomendación propuesta en Watts-Strogatz (1998) el que se denomina Hibrido (Hybrid, HYB). En esta, se parte de generar una variable B con distribución Bernoulli de una probabilidad p. Este se implementa como sigue

$$\begin{cases} Si\ X = 1\ usar\ RNE \\ Si\ X = 0\ usar\ RE \end{cases}$$

3. Muestreo Fronterizo (Frontier Sampling, FS) muestrea aristas. Fue propuesto en Ribeiro-Towsley (2010) usando paseos aleatorios. Requiere de una función para estimar y remover sesgos que el paseo aleatorio introduce.

Este es implementado por el pseudo código que se describe a continuación:

Fijar t=0

Seleccionar un conjunto de nodos $\{v_i, \dots, v_n\}$

Selecciona r una semilla v_i con probabilidad

$$P(v) = \frac{k_v}{\sum_{u \in s} k_u}$$

la arista (v_i, w) es seleccionada aleatoriamente de las de salida

Cambiar $v_i = w$

Cambiar t=t=t+1 si t<n

FS requiere que al menos un nodo cuyo grado, de salida y entrada, sea mayor que cero. Se espera que el número de ellos sea pequeño en G. FS es muy bueno en la aproximación de los grados, vea Ribeiro-Towsley (2010) y NMSE es relativamente muy pequeño. Además, se obtiene nuevos valores de C pero su desempeño no es bueno si los grados o C son pequeños.

2.5. Técnicas de exploración

El principio que les caracteriza es que en ellos se selecciona a un nodo aleatoriamente y se exploran los nodos en una vecindad del mismo. Esto brinda un marco teórico para la búsqueda de contactos de sospechosos. El posible contagiado puede ser considerado como generado por un mecanismo aleatorio y este identifica sus contactos. Los contactos identifican los suyos y se hace una cadena. Veamos los que consideramos mas flexibles

1. El llamado método de Vecindad Aleatoria del Nodo (Random Node Neighbor, RNN) se basa en seleccionar el clúster de nodos formado por el seleccionado y por todos los que se conecta con él. Este remeda la lectura de un fichero d aristas. Funciona bien cuando nos interesa estimar parámetros relacionados con los grados de salida (contactos con un infestado, por ejemplo). La distribución que se obtiene es buena para estudios de los grados de salida, pero no lo es para los grados de entrada (los infestados que tuvieron contactos con el previamente) y la estructura de la conexión de la población de vecinos generados por las aristas de salida.

2. Los paseos aleatorios son usados en el Paseo Aleatorio Simple (Random Walk, RW). Estos consisten en seleccionar aleatoriamente un nodo y simular un paseo aleatorio sobre el grafo usando una probabilidad prefijada. Es común usar $p = 0{,}15$. Se regresa al nodo raíz (por ejemplo, el contagiado seleccionado) y se recomienza el paseo. Este método puede fallar cuando se selecciona un nodo que representa un mínimo o máximo relativo o está poco conectado. Una solución es fijar un número de pasos y si no son suficientes nodos seleccionar otro individuo y repetir el proceso.

3. Una variante de RW es hacer el salto aleatoriamente. Este es llamado Saltos Aleatorios (Random Jump, RJ). En este se selecciona un camino aleatorio y se selecciona otro nodo del grafo. Este no adolece de las deficiencias de RW.

4. Otro método es el denominado Fuego Forestal (Forest Fire, FF). Este se inspira en el estudio de la evolución de grafos en el tiempo. Consiste en seleccionar aleatoriamente un nodo v, (nodo raíz) y moverse a los nodos

con que se conecta (burning outgoing links). Esto es una vez seleccionado un contagiado se seleccionan aleatoriamente algunos de los individuos con los que contactó. Si una arista (un contacto) es "quemada" el nodo con que se conecta v también se quema y prosigue el proceso recursivamente. Este modelo requiere de dos probabilidades: la de moverse hacia delante p_f y hacia atrás p_b. Ver detalles en Gjoka (2010). Veamos un algoritmo para implementarle.

Inicializar t=0,

Seleccionar v aleatoriamente

Generar aleatoriamente, de acuerdo a una distribución geométrica la variable aleatoria Z de esperanza $E(z) = \frac{p_f}{1-p_f}$

Seleccionar los nodos $w_1, \dots, w_z$ conectadso como slaidas de v que no hayan sido visitados.

Repetir el paso 3 para cada w_j, j=1,..,z.

Actualizar t=t+1

Si t<n entonces regresar al paso

Note que un nodo quemado no es visitado de nuevo y con ello se previene el caer en un ciclo. Si se "apaga" el fuego generamos otro v y seleccionamos otros nodos hasta completar la visita de n nodos raíz.

4. ALGUNOS MÉTODOS INFERENCIALES

4.1. Determinación de un grafo máximo verosímil

Consideremos un modelo en el que tenemos el conjunto de todos los posibles grafos Ω. Este incluye los grafos sin arista. Asumiendo que un mecanismo aleatorio, generado por una medida de probabilidad P, determina grafos aleatorios G_s en un grafo G poblacional. Este genera en los grafos las relaciones entre pares. Estas pueden ser medidas por una métrica

$$D: \Omega \times \Omega \to \mathbb{R}^+$$

Cada grafo $g \in \Omega$ es caracterizado por su matriz de adyacencia

$$\underline{A}_g = \left[a_{jh(t)}\right]_{n\times n};\ a_{jh(t)} = \begin{cases} 1\ si\ los\ nodos\ h\ y\ j\ están\ conectados\ en\ el\ grafo\ t \\ 0\ en\ otro\ caso \end{cases}$$

En el estudio de redes sociales Bank-Carley (1994) analizaron varios procedimientos para derivar adecuadas aproximaciones para caracterizar los modelos estocásticos que generan una red social. Usando los resultados de Mallows (1957) se derivó una medida de probabilidad definida sobre el conjunto de las posibles permutaciones y se tiene que

$$Prob(G_s|t,\sigma) = K(\sigma)exp\big(-D(G_s,G)\big),$$

donde σ es una medida de escalar y K(σ) es una medida normalizadora que no depende de t.

Si en vez del grafo real consideramos y grafo $G_0 = t \in \Omega$ que representa una medida de tendencia central de los grafos generados sobre G por $Prob(G_s|t,\sigma)$.

Tomando una muestra de m grafos generados por P. $\prod_{s=1}^{m} Prob(G_s|t,\sigma)$ es la función de verosimilitud. Su maximización es obtenible al trabajar con su logaritmo

$$log\left[\prod\nolimits_{s=1}^{m} Prob(G_s|t,\sigma)\right] = mlog\big(K(\sigma)\big) - \sigma\sum_{s=1}^{m} D(G_s,t)$$

El máximo se obtiene al minimizar el segundo término de la ecuación anterior, así que la solución del problema de optimización

$$G_0 = Arg\ min_t\left\{\sum_{s=1}^{m} D(G_s,t) = \Delta(t)\right\}$$

Este grafo es la estimación Máximo Verosímil del grafo de tendencia central. Bouza-Allende (2002) propusieron utilizar un algoritmo de Recocido Simulado para obtener su solución cuando

$$\sum_{s=1}^{m} D(G_s, t) = \sum_{s=1}^{m}\sum_{j=1}^{n}\sum_{i=1}^{n}\left|a_{jh(s)} - a_{jh(t)}\right|$$

una vez determinado el grafo máximo verosímil el epidemiólogo podrá hacer estudio sobre este y los resultados tienen la propiedad de ser lo más verosímiles obtenibles al generalizarles a la población total.

3.2. Pruebas Estadísticas para el patrón de G

La comparación de los patrones de dos grafos es realizada usando el estadístico de prueba de Kolmogorov-Smirnov D para dos muestras. En este contexto D es usado solo para medir la concordancia entre dos distribuciones. Como se sabe D = Max $\{|F_0(x) - F(x)|\}$, $x \in (x_a, x_b)$ es una variable aleatoria definida en un cierto rango de valores, F y F_0 son dos CDF empíricas de la data. Su evaluación para Scale-down y Back-in-time camping mide cuanto concuerdan la del grafo muestral con el verdadero.

Estas pruebas son muy importantes cuando consideremos el comparar el comportamiento de dos poblaciones. Por ejemplo, dos provincias. Una pregunta del epidemiólogo es si lo observado es significativamente diferente. Si los grafos determinados son aceptablemente iguales políticas particularizadas para el tratamiento post pandemia no son necesarias.

3.3. Las probabilidades de contacto

Esto es uno de los problemas cruciales en la caracterización de la expansión de contagio. Preguntas como si los riesgos de contagios son afectados por variables exógenas, como sexo, antecedentes patológicos, lugar de residencia etc., generan un marco para ser respondidas al fijar las posibilidades de contacto y establecer si hay diferencias significativas entre ellas.

Intuitivamente en una buena muestra la probabilidad de que un paseo aleatorio en G_s que comenzó en el nodo v, que visita w, debe ser similar al observado en G. Para cada nodo v en G y en G_s calculamos la probabilidad estacionaria del paseo. Se usa la norma de Frobenius para medir la diferencia entre las probabilidades de visitar (acceder).

Las epidemias podemos modelarles pensando en que son direccionales las interacciones. O sea que consideramos genera un grafo dirigido $G_d = (V, E_d)$, V es el conjunto de nodos (personas) y E_d es el conjunto de aristas direccionales (interacciones entre personal). Denotemos por $(u, v), u, v \in V$ la conexión entre el nodo u y el nodo v (arista), $k_{v_{in}}$ es el grado de entrada en el nodo v (número de aristas), $(u, v), u \in V$ en E_d, $k_{v_{out}}$ es el grado de salida del nodo v (número de aristas), $(v, w), w \in V$ en E_d.

En los grafos de salida en los que las interacciones son no-dirigidas podemos considerar que es un grafo dirigido simétrico. O sea, $\forall(u, v) \in E_d$, $(v, u) \in E_d$.

Podemos generar estos grafos simétricos s a partir de los dirigidos

Sea $G = (V, E)$ un grafo simétrico de G_d ,

$$E = \cup\{(u, v), (v, u)\}, \forall(u, v) \in E_d$$

Definamos kv como el grado del nodo v en G (numero de aristas que conectan con v)

Algunas asunciones necesarias son:

- Podemos conocer las aristas que parten de cada nodo.
- El algoritmo de selección emplea el mismo tiempo para obtener cualquier grafo muestral.

3.4. Las distribuciones

En el estudio de los grafos la distribución de los grados del nodo (NDD) caracteriza las más importantes propiedades de un grafo. Si este es dirigido un nodo posee grados de entrada y de salida. En una epidemia al tomar una persona representada por el nodo v estos representan el número de personas que se relacionan (interactúan) con esta. Una métrica para medir el comportamiento de una persona está dada a partir de tomar θ_k como la fracción de nodos cuyos grados de entrada y salida son menores o iguales a k. El error cuadrático medio normalizado (NMSE) de grado k es:

$$\text{NMSE(k)} = \frac{\sqrt{E\left(\hat{\theta}_k - \theta_k\right)^2}}{\theta_k}$$

$\hat{\theta}_k$ es una estimation de θ_k obtenida al evaluar el grafo muestral. Nos sirve para mostrar la diferencia entre la distribución de los grados del grafo muestral y el original.

Otra medida es el Coeficiente de Clustering (Clustering Coefficient, CC). Este mide el grado en que los nodos del grafo tienden a forma un clúster. Para grafos no-dirigidos este es:

$$C_u = \begin{cases} \dfrac{2|E_{v,w}|}{k_u(k_u-1)} & si\ k_u > 1 \\ 0 & en\ otro\ caso \end{cases}$$

$E_{v,w}$ es el conjunto de aristas entre los vecinos del nodo u. Su promedio en la red es (network average clustering coefficient, NACC) sobre todos los nodos del grafo es

$$\bar{C} = \frac{1}{n}\sum_{u=1}^{n} C_u$$

La distribución acumulativa de CC, denotada por γ_c, es la fracción de nodos tales que $C \leq c$. La CDF de C es:

$$\text{NMSE(c)} = \frac{\sqrt{E(\hat{\gamma}_c - \gamma_c)^2}}{\gamma_c}$$

$\hat{\gamma}_c$ es una estimación de γ_c basada en el grafo muestra.

Cuantificamos el desempeño de los algoritmos sobre diferentes propiedades del grafo mediante el error relativo (RE) el grafo muestral y el original:

$$RE = \frac{|muestra - original|}{original}$$

3.5. Análisis de las propiedades

Los métodos que implementen los epidemiólogos para trata diversos aspectos de los contagiados son evaluables a partir de estudiar las CDF de los grados de los nodos y usando NMSE para evaluar cómo se desempeñan los algoritmos de muestreo en caracterizar los grados de los nodos.

Se parte de poseer el grafo original G, por tanto, conocemos las fracciones θ_k. Para BFS y MHRW necesitamos recobrar el sub-grafo muestral $G_s = (V_s, E_s)$ a partir de los muestreados nodos. Sean V_s el conjunto de nodos en la muestra, y E_s el de las aristas. Así que

$$E_s = \cup \{(\mathrm{u}, \mathrm{v}) \in \mathrm{E}_d\}, \mathrm{u} \in V_s, \mathrm{v} \in V_s$$

G_s permite obtener la distribución.

Para FS una función es usada para estimar la CDF, ver por ejemplo [7]. Por la ley fuerte delo grandes números este estimador converge al valor real al genera una muestra grande. Las aristas en la muestra son entradas y la distribución estimada de los grados es la salida. Al aplicar FS se obtiene que $E_s = \{(u_i, v_i), i = 1, \ldots, B\}$ y el estimador de θ_k es

$$\hat{\theta}_{k_i} = \frac{1}{SB} \sum_{v_i} \frac{I\left(k_{v_i}^i \geq k^i\right)}{k_{v_i}}, i = 1, \ldots, B$$

$$S = \frac{1}{B} \sum_{v_i} \frac{1}{k_{v_i}}$$

$\hat{\theta}_{k_i}$es un estimador de θ_{k_i}, fracción de nodos cuyo grado es no mayor de k_i.

Para analizar CC comparamos el NACC obtenido por el algoritmo y el de G. para BFS y MHRW, se obtiene G_s a partir de los nodos muestreados. Entonces se calcula el promedio de los CCs de of G_s. Para FS se requiere del uso del estimador

$$\hat{C} = \frac{1}{SB} \sum_{v_i} \frac{\hat{c}(v_i)}{k_{v_i}}, i = 1, \ldots, B, \hat{c}(v_i) = \frac{2f(u_i, v_i)}{k_{v_i}\left(k_{v_i} - 1\right)}$$

La estimación de NMSE es obtenida al correr el algoritmo M veces y promediar

$$\widehat{\mathrm{NMSE}}(\mathrm{k}) = \frac{\sqrt{\frac{1}{M} \sum_{r=1}^{M} \left(\hat{\theta}_{kr} - \theta_k\right)^2}}{\theta_k}$$

$\hat{C}$es un estimador para el promedio del CC mientras que $f(u, v)$ es el número de vecinos entre u y v.

En el estudio del desempeño de los algoritmos de muestreo se analiza el ploteo de CDF y NMSE de CC.

Por su parte el estimador de FS para γ_c es:

$$\hat{\gamma}_{k_i} = \frac{1}{SB} \sum_{v_i} \frac{I(\hat{c} \leq c)}{k_{v_i}}, i = 1, \ldots, B$$

5. CONCLUSIONES

En este trabajo presentamos modelos que permiten hacer estudios en redes. Estos modelos permiten sostener investigaciones sobre el desarrollo de pandemias pues pueden ser usados para estimar parámetros como el número de contactos, densidad de los infestados etc. Las dinámicas de los procesos de infestación pueden ser caracterizados permitiendo el desarrollo de simulaciones y evaluar el efecto de políticas sanitarias en pandemias como la del COVID19.

REFERENCES

Ahn, Y., S. Han, H. Kwak, S. Moon, and H. Jeong (2007): **Analysis of Topological Characteristics of Huge Online Social Networking Services**, In Proc. of WWW.

An, J., X. Liao, et al. (2020): Clinical characteristics of the re-detectable positive RNA test, medRxiv preprint **doi: https://doi.org/10.1101/2020.03.26.20044222. (revisado Mayo, 2020)**

Banks D. and K. Carley (1994): Metric inference for social networks **J. of Classification**, 11, 121-149.

Benevenuto, F., T. Rodrigues, M. Cha, and V. Almeida (2009): , "Characterizing User Behavior in Online Social Networks," In Proc. of ACM IMC.

Bouza C. N. (2000): Muestreo adaptativo bajo contactos aleatorios. **Investigación Operacional**, 21, 38-45.

Bouza C. N. Y S. M. Allende (2002) Inferencias Sobre Grafos. [2002], **Economic Analysis Working Papers**, 1.

Bouza C. N., S. M. Allende and M. Negreiros (2015): S**ome results on sampling populations with a graph structure.**

Callaway, D. S., M.E.J. Newman, S.H. Strogatz & D.J. Watts (2000): Network robustness and fragility: Percolation on random graphs, Phys. Rev. Lett. 85, 5468–5471.

Carter, N., C. Hadlockand and D. Haughton (2008): Generating random networks from a given distribution. **Computational Statistics & Data Analysis**. 52**,** *3928-3938*

Diekmann O and J.A.P. Heesterbeek (2000): Mathematical Epidemiology of Infectious Diseases, Wiley, Chichester.

Frank, O. (1981): A survey of statistical methods for graph analysis**. In Sociological Methodology**, (S. Ferguson, N.M., D.A.T. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn and D.S. Burke (2005): Strategies for containing an emerging influenza pandemic in Southeast Asia, **Nature** 437, 209–214.

Gjoka, M. (2010): **Measurement of Online Social Networks**,U. C. Irvine PhD Thesis.

Lloyd A.L. and R.M. May (2001): Epidemiology: How viruses spread among computers and people, **Science** 292, 1316–1317

Gjoka, M., M. Kurant, C. T Butts, and A. Markopoulou (2010**): Walking in Facebook: A Case Study of Unbiased Sampling of OSNs**, In Proc. of IEEE INFOCOM, 2010.

Jin L. , Y. Chen , P. Hui , C. Ding , T. Wang , A. V. Vasilakos , B. Deng and X. Li (to be published): **Albatross Sampling: Robust and Effective Hybrid Vertex Sampling for Social Graphs**. 20XX ACM. CHDWVDL-ASREHVSSG-11.pdf (consultado Mayo 2020)

Leskovec J. and C. Faloutsos (2006): **Sampling from Large Graphs**, In Proc. of ACM SIGKDD, 2006.

Keeling M.J. and K.T.D. Eames (2006): Networks and epidemic models, **J. R. Soc. Interface** 2, 295–307

Kwak, H., C. Lee, H. Park, and S. Moon (2010): **What is Twitter, a Social Network or a News Media**? In Proc. of WWW, 2010.

Kurant, M., A. Markopoulou, P. Thiran (2010): **On the Bias of Breadth First Search (BFS) and of Other Graph Sampling Techniques**, International Teletraffic Congress, 2010.

Lee, S. H., P. -J. Kim, and H. Jeong (2006): **Statistical Properties of Sampled Networks**,Physical Review E. (consultado Abril, 2020)

Longini, I.M., A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D.A.T. Cummings and M.E. Halloran (2005): Containing pandemic influenza at the source, **Science** 309, 1083–1087

Mallows, C. (1957): Non null ranking models I. **Biometrika** 44, 114-130.

Meyers, L.A. B. Pourbohloul, M.E.J. Newman, D.M. Skowronski & R.C. Brunham (2005): Network theory and SARS: Predicting outbreak diversity. **J. Theor. Biol**. 232, 71–81.

Mislove, A., M. Marcon, K. P. Gummadi, P. Druschel, and B. Bhattacharjee (2007): **Measurement and Analysis of Online Social Networks**, In Proc. of ACM IMC, 2007.

Newman M.E.J. (2002): The spread of epidemic disease on networks, **Phys. Rev. E** 66, 016128.

Ribeiro B. and D. Towsley (2010): **Estimating and Sampling Graphs with Multidimensional Random Walk**s, In Proc. of ACM IMC.

Sala, A. L. Cao, C. Wilson, R. Zablit, H. Zheng, and B. Y. Zhao (2010): **Measurement-calibrated Graphs Models for Social Network Experiments,** In Proc. of WWW, 2010.

Thompson, S. K. (1990): Adaptative cluster sampling. **J. Amer. Stt. Ass**. 67, 224-227.

Wang, T., Y. Chen, Z. Zhang, P. Sun, B. Deng, and X. Li (2010): **Unbiased Sampling in Directed Social Graphs**, In ACM SIGCOMM Computer Communication Review, 40,401-402.

Watts D. J. and S. Strogatz (1998): Collective Dynamics of 'Small-World' Networks, **Nature** 393, 440-442.

Wilson, C., B. Boe, A. Sala, K. P. N. Puttaswamy, and B. Y. Zhao (2009): **User Interactions in Social Networks and their Implications**, In Proc. of ACM EuroSys, 2009.

Capítulo 4 pp 47-51

VINCULACIÓN DEL PROGRAMA R CON GOOGLEMAP PARA ANÁLISIS ESPACIALES

José A. Betancourt Bethencourt[1]
[1]Universidad de Ciencias Médicas Carlos J. Finlay, Cuba

RESUMEN

La epidemiología espacial se utiliza para describir, cuantificar y explicar las variaciones geográficas de las enfermedades; para evaluar la relación entre la incidencia de las mismas y posibles factores de riesgo asó como para identificar los conglomerados geográficos de las enfermedades.
Se simuló la ubicación cartográfica de la presentación de casos de COVID-19 en el área de salud Tula Aguilera del municipio Camagüey durante enero de 2022, para poder visualizar estos en mapas con googlemap se preparó un script en Rmarkdown (RMD). Se presentó un mapa de calor (heatmap) y un mapa de incidencia (bublemap). El objetivo del presente trabajo estriba en entrenar de manera interdisciplinaria a estudiantes y residentes de la provincia de Camagüey en la vinculación del programa R con googlemap para análisis espaciales para apoyar la toma de decisiones y las acciones de intervención ante brotes epidémicos. Se concluye que fue posible hacer mapas de riesgos y de incidencia y realizar su interpretación para entrenar estudiantes y residentes.

PALABRAS CLAVE: epidemiología, mapas, R, aprendizaje, googlemap

ABSTRACT

Spatial epidemiology is used to describe, quantify, and explain geographic variations in disease; to evaluate the relationship between their incidence and possible risk factors as well as to identify the geographic conglomerates of the diseases. The cartographic location of the presentation of COVID-19 cases in the Camagüey municipality during January 2022 was simulated, in order to visualize these on maps with googlemap; an Rmarkdown (RMD) script was prepared. A heat map (heatmap) and an incidence map (bublemap) were presented. The objective is to train students and residents in linking the R program with googlemap for spatial analysis. It was possible to make risk and incidence maps and perform their interpretation to train students and residents.

KEY WORDS: epidemiology, maps, R, learning, googlemap

1. INTRODUCCIÓN

La epidemiología espacial se utiliza para describir, cuantificar y explicar las variaciones geográficas de las enfermedades; para evaluar la relación entre la incidencia de las mismas y posibles factores de riesgo, así como para identificar los conglomerados geográficos de las enfermedades de manera sucinta y clara. Con los mapas entonces se hace factible planificar y dirigir una respuesta rápidamente y así planificar y monitorear programas de intervención y erradicación en la preparación para emergencias (Palú Orozco et al., 2021)

[1] e-mail: betanster@gmail.com

El uso de técnicas de análisis espacial en los estudios epidemiológicos ha tenido un rápido crecimiento en las últimas décadas porque permiten incluir en los análisis la variabilidad espacial mapas coropléticos, pruebas de hipótesis y la identificación de clústeres de enfermedades.(Valbuena-Garcia and Rodríguez-Villamizar, 2018a)

El objetivo del presente trabajo estriba en entrenar a estudiantes y residentes de la provincia de Camagüey de manera interdisciplinaria en la vinculación del programa R con googlemap para análisis espaciales para apoyar la toma de decisiones y las acciones de intervención.

2. MÉTODOS

En el presente estudio cuantitativo aplicado, se simuló la ubicación cartográfica de la presentación de casos de COVID-19 en el área Tula Aguilera del municipio Camagüey durante enero de 2022, para poder visualizar estos en mapas con googlemap se preparó un script en Rmarkdown (RMD). Se presentó un mapa de calor (heatmap) y un mapa de incidencia (bublemap).

3. RESULTADOS Y DISCUSIÓN

El programa R(R_Core_Team, 2021) es libre y de fácil programación. Con el RMD utilizado se hizo posible "llamar" a Googlemap y ubicar los puntos cartográficos de presentación de casos para ver tanto el riesgo (Figura 1) como la incidencia puntual (figura 2). Este RMD puede utilizarse para valorar no solo los casos de COVID-19 sino casos de cualquier enfermedad transmisible.

En diversos estudios se detalla cómo estas herramientas geográficas son adecuadas para analizar la in formación existente y mejorar las acciones de control de brotes epidémicos (Moraga, 2017);(Lacabana, 2019, Valbuena-Garcia and Rodríguez-Villamizar, 2018b);(Aswi et al., 2019).

La relación de los puntos calientes con los ejes viales señalaría que la movilidad e interacciones socioeconómicas entre localidades cercanas, son factores que intervienen en la ocurrencia de las incidencias. Estas evidencias constituyen una herramienta destinada a facilitar la interpretación de la realidad y continuar una línea de investigación orientada hacia la implementación de distintas estrategias de intervención.(Palú Orozco et al., 2021)

En este contexto, dentro de los análisis de densidad, se han desarrollado los estudios de los mapas de calor. Es así como mediante estos procesos es posible mostrar las áreas donde se produce la mayor densidad de un evento. Por este motivo las técnicas basadas en análisis de mapas de calor se configuran como muy útiles para el tratamiento y la prevención de diversas problemáticas que afectan a la calidad de vida de las personas. Particularmente, los análisis de densidad no son un análisis de mapa de calor propiamente dicho, sino un método de interpolación que estima la probabilidad de que ocurra cierto evento en un área determinada, quedando, de este modo, identificadas las áreas de mayor riesgo.(Hotjar-tools, 2021, Prado-Ortega and Grunauer-Robalino, 2020)

Se han utilizado los mapas de densidad para evaluar la incidencia de tuberculosis en España (Dominkovics et al., 2011) y para el análisis del SIDA en Brasil (Sousa et al., 2021). Se aboga por el uso interactivo del R en la creación de mapas de calor,(Gu and Hübschmann, 2022). Se ha valorado con análisis espaciales el elevado riesgo de aerosoles de Sarcov-2 para especialistas mayores en otorrinolaringología y aumento de cargas virales (Ruthberg et al., 2020)

Coincidimos con (Mejía, 2019) en que el principal reto para el avance de la aplicación de los SIG es lograr que los tomadores de decisiones conozcan sobre los múltiples beneficios que se pueden alcanzar al analizar la

información generada desde los establecimientos de salud, aprovechando la tecnología espacial. Una de las principales apuestas a futuro debe ser la capacitación del personal técnico de salud.

El autor está plenamente de acuerdo con (Hernández Galvez et al., 2021) en que la enfermedad de COVID-19 ha impuesto en la educación médica un pensamiento de cambio y transformación del proceso de enseñanza y aprendizaje tradicional, al cual debe sumarse y engranarse un nuevo escenario virtual con el uso óptimo de las tecnologías de la información y las comunicaciones. El reto está en identificar las acciones positivas que permitan un desarrollo con agilidad y destreza, pues el futuro de la educación médica será diferente, por lo que hay que asumir con responsabilidad la acción de formar a los estudiantes como profesionales competentes en su desempeño.

4. CONCLUSIÓN

Se concluye que la herramienta presentada es adecuada, en la misma se vincula al programa R con googlemap para análisis espaciales que facilitan la toma de decisiones y las acciones de intervención ante los brotes epidémicos.

REFERENCIAS

ASWI, A., CRAMB, S., MORAGA, P. & MENGERSEN, K. 2019. Bayesian spatial and spatio-temporal approaches to modelling dengue fever: a systematic review. *Epidemiology & Infection* [Online], 147. Available: https://www.cambridge.org/core/journals/epidemiology-and-infection/article/bayesian-spatial-and-spatiotemporal-approaches-to-modelling-dengue-fever-a-systematic-review/172B331C11CDDA3C0A14C667708C248F.

DOMINKOVICS, P., GRANELL, C., PÉREZ-NAVARRO, A., CASALS, M., ORCAU, À. & CAYLÀ, J. A. 2011. Development of spatial density maps based on geoprocessing web services: application to tuberculosis incidence in Barcelona, Spain. *International journal of health geographics,* 10**,** 1-14.

GU, Z. & HÜBSCHMANN, D. 2022. Make interactive complex heatmaps in R. *Bioinformatics* [Online], 38. Available: https://academic.oup.com/bioinformatics/article/38/5/1460/6448211?login=false.

HERNÁNDEZ GALVEZ, Y., LÓPEZ ARBOLAY, O. & FERNÁNDEZ OLIVA, B. 2021. Nueva realidad en la educación médica por la COVID-19. *Educación Médica Superior,* 35.

HOTJAR-TOOLS. 2021. Que es un mapa de calor. Available: https://www.hotjar.com/heatmaps/.

LACABANA, P. 2019. Sistemas de información geográfica para la toma de decisiones. El dengue en el partido de Quilmes. *Ambiente y desarrollo sustentable: miradas diversas* [Online]. Available: https://ridaa.unq.edu.ar/bitstream/handle/20.500.11807/289/PGDeBook_ambiente_2017_001.pdf?sequence=1&isAllowed=y#page=108.

MEJÍA, R. 2019. Sistemas de información geográfica y su aporte a la salud pública en El Salvador. *Alerta, Revista científica del Instituto Nacional de Salud,* 2**,** 71-74.

MORAGA, P. 2017. SpatialEpiApp: A Shiny web application for the analysis of spatial and spatio-temporal disease data. *Spatial and Spatio-Temporal Epidemiology* [Online], 23. Available: https://eprints.lancs.ac.uk/id/eprint/87484/1/1_s2.0_S187758451730062X_main.pdf.

PALÚ OROZCO, A., TEXIDOR GARZÓN, M. C., PORTUONDO PUJOL, C., MIRANDA REYES, S. C. & MANET LAHERA, L. R. 2021. Telepidemiología en el enfrentamiento a la COVID-19 en la provincia Santiago de Cuba. *Revista Cubana de Salud Pública,* 47.

PRADO-ORTEGA, M. X. & GRUNAUER-ROBALINO, G. R. 2020. SALUD PÚBLICA: DETECCIÓN DE CONCENTRACIÓN DE PERSONAS APLICANDO BIG DATA PARA EVITAR BROTES EPIDEMIOLÓGICOS COVID-19. *Identidad Bolivariana,* 4**,** 5-19.

R_CORE_TEAM. 2021. A language and environment for statistical computing. R Foundation for Statistical Computing. . Available: https://www.r-project.org/.

RUTHBERG, J. S., QUERESHY, H. A., JELLA, T. K., KOCHARYAN, A., D'ANZA, B., MARONIAN, N. & OTTESON, T. D. 2020. Geospatial analysis of COVID-19 and otolaryngologists above age 60. *American journal of otolaryngology,* 41**,** 102514.

SOUSA, L. C., SILVA, T. C., FERREIRA, T. F. & CALDAS, A. D. J. M. 2021. Spatial analysis of AIDS in the state of Maranhão: an ecological study 2011-2018. *Revista Brasileira de Enfermagem,* 75.

VALBUENA-GARCIA, A. M. & RODRÍGUEZ-VILLAMIZAR, L. A. 2018a. Análisis espacial en epidemiología: revisión de métodos. *Revista de la Universidad Industrial de Santander. Salud,* 50**,** 358-365.

VALBUENA-GARCIA, A. M. & RODRÍGUEZ-VILLAMIZAR, L. A. 2018b. Análisis espacial en epidemiología: revisión de métodos. *Revista de la Universidad Industrial de Santander. Salud* [Online], 50. Available: http://www.scielo.org.co/scielo.php?script=sci_arttext&pid=S0121-08072018000400358.

ANEXOS

Figura 1. Mapa de riesgo

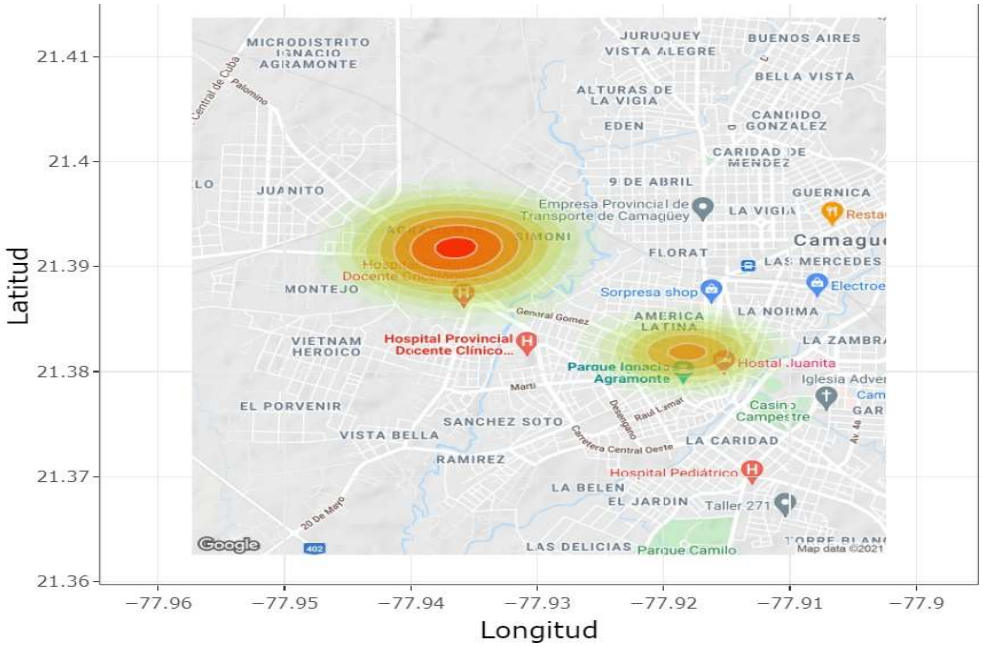
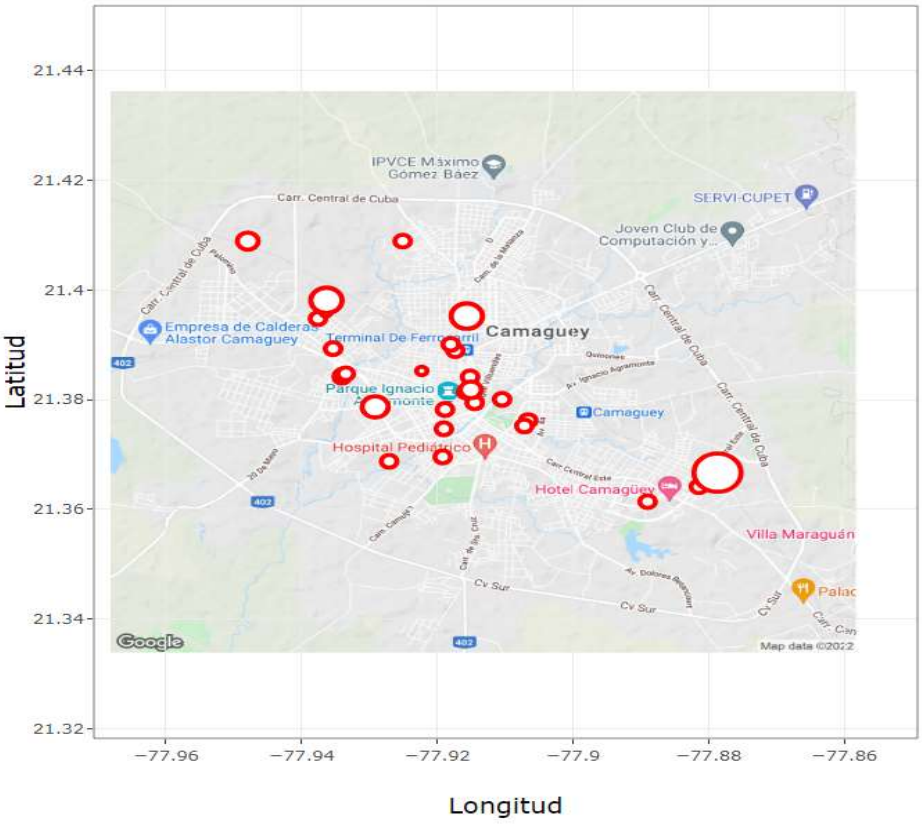


Figura.2 Incidencia de casos, el tamaño refleja el número de casos


21.44
21.42
21.4
21.38
21.36
21.34
21.32
Latitud
−77.96
−77.94
−77.92
−77.9
−77.88
−77.86
Longitud
IPVCE Máximo Gómez Báez
Carr. Central de Cuba
SERVI-CUPET
Joven Club de Computación y...
Empresa de Calderas Alastor Camaguey
Terminal De Ferrocarril
Camaguey
Parque Ignacio Agramonte
Hospital Pediátrico
Hotel Camagüey
Villa Maraguán
Google
Map data ©2022

Capítulo 5 pp 53-85

ALGUNOS ELEMENTOS SOBRE LAS CURVAS ROC: TEORÍA Y HERRAMIENTAS

Carlos N. Bouza-Herrera[1], Pablo Otoniel Juárez Moreno[2] y Octaviano Juárez Romero[2]

[1]Universidad de La Habana, Cuba
[2]Universidad Autónoma de Guerrero, México

ABSTRACT

These monograph aims to give a vision of the problematic of the ROC curves. The different elements on them are disperse in the literature and there are not many materials dealing with the mathematical, interpretation of graphs, measures etc., in a unified way. Examples are presented and their interpretation are discussed. It is expected this material will be useful for analyzing data in the areas where ROC curves are of common use.

Keywords: ROC curves, mathematic modeling, graphs, statistical measures, Covid.

RESUMEN

En esta monografía se trata de brindar una visión de la problemática de las curvas ROC. Los elementos de estas están dispersos en la literatura y no abundan unificadamente materiales donde se presenten los problemas matemáticos, de interpretación de gráficos, medidas etc. Se presentan ejemplos y se discute como interpretarles. Se espera sea útil para su uso en el análisis de datos en las áreas donde las curvas ROC son utilizadas.

Palabras clave: curvas ROC, modelación matemática, gráficos, medidas estadísticas, Covid.

1. INTRODUCCIÓN

Las curvas ROC son de amplio uso para evaluar el desempeño de métodos clasificatorios. Estos tratan de identificar a que tipo de eventos pertenecen las observaciones (individuos, unidades instancias etc.). Para producir tales curvas en general se requiere de una muestra de observaciones es utilizada para determinar una "regla de oro" contra la que comparar el desempeño de clasificadores. Esto puede ser resuelto usando métodos más complejo como señalan Krzanowski y Hand (2009), Fawcet (2003), Hui y Zhou (1998) y Peizhou et al. (2017). A partir de una muestra la curva ROC resultante determina una función escalonada en el plano. En el caso continuo la curva ROC representa una continua. En la mayoría de los casos esta no es conocida y es aproximada por una función escalonada obtenida a partir de la distribución empírica. Esta puede ser suavizada usando algún método numérico o estadístico. Discusiones sobre esto se pueden obtener en Zou, et al. (1997) Lloyd (1998), Metz (1978), Wan y Zhang (2007), Drummond y Holte (2004), Qiu y Le (2001).

En general consideramos que las curvas ROC son determinadas a partir de definir una clase como positiva y otra como negativa. Una instancia es clasificada en una de esas clases a partir de un valor llamado score. Una muestra se observa y se obtiene un conjunto de observaciones y el método de clasificación determina una distribución de los scores para cada clase. El decisor establece el rango de los scores que permite decidir en qué

[1] E-mail: bouza@matcom.uh.cu

clase ubicar una observación. El problema que le sigue es determinar una medida del comportamiento del método de clasificación.

Lo más popular es su uso en la evaluación del desempeño de tests diagnósticos en medicina para establecer si un paciente es positivo (sano, contaminado etc.) o negativo (enfermo, no contaminado etc.) Fuentes importantes de elementos y ejemplos pueden obtenerse en Cerda y Cifuentes (2012), Martínez-Camblor (2007) entre otros. En estudios forenses las curvas ROC son usadas para establecer la concordancia o no de muestras de ADN y otras pruebas identificadoras. En informática estas son una herramientita para filtrar mensajes como spam o no, por ejemplo, y dejarles entrar en los sistemas. En Inteligencia Artificial es útil para evaluar el comportamiento de la predicción de Redes Neuronales Artificiales, por ejemplo, vea Fawcett (2003, 2006) para una amplia discusión sobre su uso de minería de datos, aprendizaje por computadora etc.

En la sección 2 se presentan discusiones sobre la teoría matemática que sustenta el análisis de las curvas ROC. Los intríngulis de los enfoques son desmontados al considerar el no paramétrico, el paramétrico y el semi-paramétrico. Sobre estos aspectos se puede abundar en Hsieh y Turnbull (1996), Krzanowski y Hand (2009), Fawcet (2003), Gang, et al . (1999), Ertugrul et al. (2012), Wan y Zhang (2007), Cai y. Moskowitz (2004), Pardo y Franco-Pereira (2017).

La sección 3 se dedica a discutir como otras herramientas permiten complementar los análisis más tradicionales usados en estudios basados en curvas ROC. Véase Albert et al. (2014), Lloyd y Yong (1999), Macaskill et al. (2010), Drummond y Holte. (2006).

En la sección 4 se discuten algunos ejemplos de cómo aplicar los instrumentos en problemas prácticos de aprendizaje automatizado, clasificación de tumores, el uso de criterios de costo y la identificación de la contaminación por verter desechos sólidos.

Finalmente, la sección 5 presenta algunos comentarios sobre softwares para llevar a cabo estudios basados en curvas ROC. Sobre este aspecto se puede ampliar consultando Metz (2011), Carleos et al. (2010), Dorfman (2006),), Krzanowski y Hand (2009), Fawcet (2003) entre otros.

2. LA CURVA ROC

2.1. Sus elementos de base

Es común en estudios diagnósticos que un marcador continuo sea tratado primariamente usando la Curva Característica de la Operación (Receiver Operating Characteristic, ROC). Este problema es ampliamente disctuido en Martínez-Camblor, (2007), Metz (1978). Ella es obtenida al plotear la probabilidad de que el marcador esté por encima de un cierto umbral "c" de los decesos (sensibilidad) contra la probabilidad de que esté por debajo del umbral para sujetos sanos (1-especificidad). En términos más simples: se tiene la fracción de verdaderos positivos (TPF) y se plotea versus la fracción de falsos positivos (FPF). Esta está dada por la función:

$$ROC(c) = \{FCP(c), TPF(c); c \in]-\infty, \infty[\}$$

Aunque existen varios índices para medir la eficacia de un sistema diagnóstico la más usada es el área debajo de la curva ROC(AUC). Esta no depende de la unidad de medición, sumarizando la sensibilidad y la especificidad sobre el rango de los umbrales. AUC está siempre entre 0,5 y 1. Por ello es comúnmente usada para comparar el desempeño de marcadores. La curva ROC provee una descripción de la separación entre las distribuciones de positivos y negativos sin requerir de hipótesis probabilísticas. Vea a continuación una gráfica que ilustra como se soportan las decisiones en un problema médico. Los decesos y no-decesos tienen distribuciones diferentes que se sobrelapan. Al observar un score no se sabe de qué distribución lo generó. Hay

dudas sobre ello, pero en el primer paso se debe decidir sobre lo verosímil de aceptar una distribución como adecuada. Habrá una para los scores de los casos positivos y otras para la de los negativos. Una vez fijadas se debe establecer un punto partir del cual se acepta que la observación es generada por una de las usadas. Al comprar un score con este se decide por lo que sea más verosímil a la luz de estas distribuciones. Vea un ejemplo la figura siguiente que ilustra el problema de decidir si el valor del score sugiere que el paciente morirá o no. Las distribuciones de los decesos son diferentes de la de los no decesos. Ante tres puntos de corte alternativos el médico deberá decir cual usar.

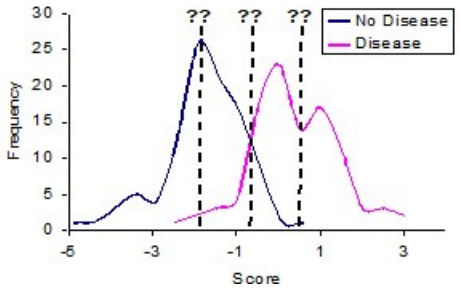


Ilustración 1: Distribuciones alternativas.

Para construir una curva ROC es necesario conocer la pertenencia a las clases de cada observación (instancia). El decisor determinará como obtenerla. Por ejemplo, en medicina se conoce quienes son considerados positivos y quienes no. En otros casos se deberá hacer algunas manipulaciones.

En el caso binario se mide la pertenencia a una clase. Conociendo la pertenencia a las clases se construye ROC para el modelo en cuestión. En el caso de varios marcadores aquel cuya ROC tiene mayor AUC será preferido. Por ejemplo, en casos médicos se tienen usualmente las dos clases "ser positivo" y "ser negativo". A partir de muestras se evalúan medidas que evalúen la exactitud del clasificador. Es común que se computen la tasa real de positivos (true positive rate, TPR) y la de los falsos positivos (false positive rate, FPR). Un estudio usará el resultado de plotear TPR versus FPR usando varios umbrales para predecir resultados y evaluar el comportamiento del fenómeno estudiado.

Otra medida útil es la distancia vertical máxima entre la curva ROC y la linea de 45º. Esta línea indica cuán lejos está la curva de la de una prueba no informativa. Esta distancia estará cerca de cero si la prueba es muy no informativa y cerca de 1 si es muy informativa. Pensando en comparar distribuciones pueden usarse las pruebas de Kolmogorov-Smirnov Anderson-Darling o la de Neyman-Watson. Varios estudios recomiendan usar Anderson-Darling por aparecer con mayor potencia. Pruebas particulares para comparar curvas ROC son dadas por Venkatraman (2000) y Martínez-Camblor, Carleos y Corral (2011).

Los métodos más usados son los basados en la distribución empírica o en asumir la bi-normalidad. Los métodos necesarios para hacer estudios basado en ROC están incluidos en la mayor parte de los paquetes estadísticos.

2. 2. Una visita a la teoría.

En el estudio de las curvas ROC es común utilizar una curva ROC empírica al procesar una muestra de instancias. Esta puede ser generada por una curva ROC real pero desconocida que pertenece a una cierta familia de funciones. La curva ROC estimada se espera se aproxime a la real. Esta problemática es discutida en la bibliografía referenciada. La conceptualización del modelar discutida en esta sub-sección es derivada de los estudios presentados en Fawcett (2006), Macaskill et al. (2010), Pardo y Franco-Pereira (2017), Pepe (2000).

Considere que el investigador obtiene una base de datos independientes $\{D_j, \pi_j; j = 1, \dots, n\}$. D_j es un score producido por el clasificador y

$$\pi_j = Prob\{j\ es\ positivo\}$$

Al observar un dato este se caracteriza mediante la variable

$$S_j = \begin{cases} 1\ si\ j\ es\ un\ verdadero\ positivo \\ 0\ en\ otro\ caso \end{cases}$$

Esta es un Bernoulli con parámetro π_j. Si las variables son independientes y son condicionalmente independientes de D_j. Para un valor el umbral (punto de corte) "c" la tasa esperada de verdaderos positivos es

$$E(TPR(c)) = E\left(\frac{1}{\sum_{i=1}^{n} S_i}\sum_{i=1}^{n} S_i I(X_i > c)\middle|\sum_{i=1}^{n} S_i > 0\right)$$

Esta esperanza es monótona decreciente respecto a c. Dada la independencia, como

$$\lim_{n\to\infty}\left\{Prob\left(\sum_{i=1}^{n} S_i > 0\right)\right\} \cong 1 - \lim_{n\to\infty}\left\{\prod_{i=1}^{n}(1-\pi_i)\right\} \geq 1 - \lim_{n\to\infty}\left(\min_{\pi_i \neq 0}\pi_i\right)^n = 1$$

La esperanza es aproximada, para un k fijo, a partir de que

$$E\left(\frac{S_k}{\sum_{i=1}^{n} S_i}\middle|\sum_{i=1}^{n} S_i > 0\right) \cong \pi_k E\left(\frac{1}{1-\sum_{i=1}^{n} S_i}\right)$$

De ahí se deriva que, usando $F_{t_k}(t)$ como la Función Acumulativa de Distribución (CDF) de $\sum_{i\neq k}^{n} S_i$

$$ETPR(c) \cong E\left(\sum_{i=1}^{n}\pi_i\vartheta_i I(X_i > c)\right);\ \vartheta_k = \left(1-\sum_{i\neq k}^{n} S_i\right)^{-1} = \frac{\sum_{t=0}^{n-2} F_{t_k}(t)}{(t+1)(t+2)} + \frac{1}{n}$$

Para los verdaderos negativos la variable Bernoulli se denota

$$R_j = \begin{cases} 1\ si\ j\ es\ un\ verdadero\ negativo \\ 0\ en\ otro\ caso \end{cases}$$

Medidas similares a las deducidas para los verdaderos positivos aparecen a continuación. La respuesta esperada de falsos positivos del clasificador es

$$EFPR(c) = E\left(\frac{1}{\sum_{i=1}^{n} R_i}\sum_{i=1}^{n} R_i I(X_i > a)\middle|\sum_{i=1}^{n} R_i > 0\right)$$

Otra medida es la esperanza

$$ETDR(c) = E\left(\frac{\sum_{i=1}^{n} S_i I(X_i > c)}{\sum_{i=1}^{n} I(X_i > c)}\right)$$

A partir de las propuestas hechas ETPR(c) y EFPR(c) pueden ser adaptadas para hacer el usual análisis que se realiza usando ROC. Para ello se hace variar el umbral. Una sucesión $\{c_1, \dots, c_m\}$ genera las sucesiones {ETPR(c_i)} and {EFPR(c_i) }. Note que el par {ETPR (c_i), EFPR (c_i) } es la representación de un clasificador

en el espacio determinado por las ROC. El grafo determinado por estos pares es una función escalera que representa la curva ROC (EROC). La curva EROC permite analizar la habilidad promedio de un clasificador, al rankear las instancias positivas respecto a la negativa, dada las incertidumbres dadas por la referenciación estándar. Así que la comparación de clasificadores es realizable simplemente al analizar el comportamiento de EROC. Así que vale calcular el área debajo de la curva EROC (EAUC). Dada la existencia de incertidumbre, no se puede identificar EAUC con la probabilidad de que el clasificador rankee una instancia positiva, obtenida aleatoriamente, por encima de una seleccionada similarmente siendo esta negativa. Sin embargo, EAUC provee una medida para establecer, en promedio, cuan grande debe ser la probabilidad de rankear correctamente instancias observadas aleatoriamente de las clases.

Como se ha visto la curva ROC es determinada por plotear las tasas de falsos positivos versus las tasas de verdaderos positivos para varios umbrales, puntos de corte, para los resultados de una prueba diagnóstico. Por su parte, EROC permite medir la incertidumbre que se tendrá a partir de la curva ROC construida. Este análisis es de común uso en la determinación de la exactitud de las pruebas diagnósticas.

Comúnmente se sumariza la exactitud a partir del área bajo la curva ROC (AUC). Ella está entre 0 y 1. Grandes valores sugieren una mayor exactitud.

El resultado de una prueba diagnóstica puede ser binario, ordinal, o continuo. Por ejemplo, la determinación del éxito o no de una vacuna es binaria, el rankeo de la eficacia de alternativas es ordinal y la medición del comportamiento de biomarcadores de problemas cardiacos y cancerígenos son de tipo continuo. A partir de la naturaleza de los resultados se selecciona el método idóneo para estimar las curvas ROC y el AUC.

2.3. El ROC no paramétrico

El análisis ROC comúnmente es no-paramétrico pues este no requiere de hacer asunciones sobre la distribución de los resultados de la prueba diagnóstica. Su popularidad se basa en que no hace uso de hipótesis sobre la CDF de las medidas obtenidas de las pruebas diagnósticas. Los resultados son obtenidos al fijar un valor umbral "c" para catalogar los verdaderos y falsos positivos. La tasa es calculada a partir del conteo de los casos en cada categoría.

El estimador natural de la curva ROC en este contexto es sustituir en la ecuación la CDF la CDF-empírica, Cerda y Cifuentes (2012), Drummond y Holte (2004), Ertugrul (2012), Fawcett (2006), Hsieh y Turnbull (1996). La curva ROC resultante no es suave.

El estimador empírico de la curva ROC parte de su definición aplicada a los datos observados. Sean D_0 y D_1 las poblaciones de no positivos y positivos, respectivamente, y c un valor predeterminado. Las correspondientes tasas son las probabilidades

$$TPF(c) = Prob\{Y \geq c|D_1\},\ FPF(c) = Prob\{Y \geq c|D_0\}$$

En teoría estas no son sino

$$\alpha PF = 1 - F_\alpha(c|D_h);\ \alpha = T, F; h = 1{,}0$$

En la práctica se calculan las tasas de verdaderos y falsos positivos, a partir de las muestras $s_h;\ |s_h| = n_h, h = 0{,}1$, obteniendo las estimaciones de las tasas teóricas mediante las proporciones (probabilidades empíricas)

$\widehat{TPF}(c) = \frac{1}{n_1}\sum_{i=1}^{n_1} I_i(Y_{D_1} \geq c), \widehat{FPF}(c) = \frac{1}{n_0}\sum_{i=1}^{n_0} I_i(Y_{D_0} \geq c)$

La curva ROC empírica es el grafo que resulta de plotear $\{\widehat{TPF}(c), \widehat{FPF}(c)\}$. Esta es una aproximación de {TPF (c), FPF (c)} donde c ∈ (−∞, ∞). La ROC empírica es entonces

$$\widehat{ROC}(c) = \widehat{TPF}\left(\widehat{FPF}(c)^{-1}\right), 1 \leq t \leq 1$$

El problema de las k-muestras es clásico dentro de la estadística no-paramétrica. Existen numerosas herramientas que permiten docimar la igualdad para muestras independientes. Vea, Martínez-Camblor, Carleos, and Corral (2011) como referencia. Esto provee de un marco adecuado para hacer inferencias estadísticas al alizar k curvas ROC.

Otros estimadores han sido desarrollados dentro del marco de la estadística no paramétrica. Vea por ejemplo Zou, Hall y Shapiro (1997) y Lloyd (1998). Ellos trataron con el problema de suavizar la curva obtenida. En particular Qiu y Le (2001) propusieron un efectivo método usando técnicas de suavizamiento local.

Al pensar en el AUC de la curva ROC no paramétrica se tiene que usando la regla trapezoidal

$$\widehat{AUC} = \frac{1}{n_1 n_0}\sum_{j=1}^{n_0} \psi\left(Y_{i1}, Y_{j0}\right), \qquad \psi\left(Y_{i1}, Y_{j0}\right) = \begin{cases} 1 \; si \; Y_{i1} < Y_{j0} \\ \frac{1}{2} \; si \; Y_{i1} = Y_{j0} \\ 0 \; si \; Y_{i1} > Y_{j0} \end{cases}$$

es el estimador. En la fórmula se toman

$$Y_{j1} = resultado\ de\ la\ prueba\ diagnóstica\ para\ los\ decesos$$

$$Y_{jo} = resultad\ de\ la\ prueba\ diagnóstica\ para\ los\ no - decesos$$

A partir de las muestras se calculan

$n_0^{=}(Y) = número\ de\ verdaderos\ negativos\ con\ valor\ \ igual\ a\ Y,$

$$n_0^{<}(Y) = número\ de\ verdaderos\ negativos\ con\ valor\ menor\ que\ Y$$

$$n_1^{=}(Y) = número\ de\ verdaderos\ positivos\ con\ valor\ igual\ a\ \ Y$$

$$n_1^{>}(Y) = número\ de\ verdaderos\ positivos\ con\ valor\ mayor\ que\ \ Y$$

$$Q_1 = \frac{1}{n_0 n_1^2}\sum_{Y} n_0^{=}(Y)\left(n_1^{>}(Y)^2 + n_1^{=}(Y)n_1^{>}(Y) + \frac{n_1^{=}(Y)^2}{3}\right)$$

$$Q_0 = \frac{1}{n_1 n_0^2}\sum_{Y} n_0^{<}(Y)\left(n_0^{<}(Y)^2 + n_0^{=}(Y)n_0^{<}(Y) + \frac{n_0^{=}(Y)^2}{3}\right)$$

La estimación de la varianza de $\widehat{AUC}$ es derivada de las estructuras de los estadísticos lineales de rango que son del tipo Mann-Whitney. Esta es obtenida al aplicar el método delta siendo

$$\hat{V}\left(\widehat{AUC}\right) = \frac{\widehat{AUC}\left(1 - \widehat{AUC}\right) + (n_1 - 1)\left(Q_1 - \widehat{AUC}^2\right) + (n_0 - 1)\left(Q_0 - \widehat{AUC}^2\right)}{n_0 n_1}$$

Así que, las pruebas de hipótesis se desarrollan usando las correspondientes a las del tipo Mann-Whitney.

Otra forma de usar métodos no paramétricos es considerar

$$R_{n_1,n_0}(t) = \hat{F}_{1-\hat{F}_{n_1,n_0}(D_P)}(t)$$

y usar su distribución asintótica

$$R(t) = 1 - F_P\left(F_N^{-1}(1-t)\right) = F_{1-F_N(D_P)}(t)$$

F_P es la CDF de los positivos y F_N la de los negativos.

Note que de la definición de $\hat{R}_{m,n}(t)$Eq. (1) se tiene que la curva ROC es convexa.

En los estudios de las curvas ROC poseemos cierta información sobre F_N y F_P y no sobre R(t), así que la distribución de $\hat{R}_{m,n}(t)$ no va a depender de las particulares CDF´s ,F_N y F_P , sino de la expresión final de R(t) y su primera derivada, $r(t) = \frac{\partial R(t)}{\partial t}$. O sea que, conociendo la expresión analítica de la curva ROC, no hace falta conocer las CDF`s de las que se origina. Como se deduce, hay una relación directa entre la curva ROC y la distribución empírica. Véase, Molanes-López y Cao (2008) para una discusión extensa de este hecho. Esas curvas son la distribución de los individuos (instancias) positivos cuando la de los negativos es uniforme.

Se considera que tanto $F_P(t)$ como $F_N(t)$ poseesn funiones de densiad contiunas $f_P(t)$ y $f(t)$, respectivamente y que ,$\frac{f_P\left(F_N^{-1}(t)\right)}{f_N\left(F_N^{-1}(t)\right)}$ está acotada en cualquier subintervalo $(a,b) \subset (0,1)$ y que $\lim_{\min\{n_1,n_0\}\to\infty}(n_0/n_1) = \lambda$(a, b) . En este contexto Hsieh y Turnbull (1996) obtuvieron un modelo probabilístico complejo al establecer que existe un espacio de probabilidad sobre el que son definibles sucesiones que son independientes versiones de los puentes of Brownianos

$$\left\{B_1^{(n_1)}(t)\right\}_{\{0\le 1\le t\}}, \left\{B_2^{(n_0)}(t)\right\}_{\{0\le 1\le t\}}$$

Tales que

$$\sqrt{n}(R_{n_1,n_o}(t) - R(t)) =_{cs} \sqrt{\lambda}\frac{\partial R(t)}{\partial t}\left[B_1^{(n_1)}(1-t)\right] + B_2^{(n_0)}(t)\left(1 - R(t)\right) + o(1)$$

Estimaciones, donde la curva ROC es suavizada, son obtenibles a partir del método propuesto por Qiu y Le (2001). Este se basa en técnicas de suavizamiento local.

Cuando la data proviene de experimentos pareados puede usarse la sugerencia de Venkatraman y Begg (1996). Estos desarrollaron una prueba del tipo permutación en las que el re-muestreo que garantiza que la hipótesis de intercambiabilidad sea satisfecha. Esta se basa en que se garantice la existencia de alguna transformación monótona adecuada. En general se emplea algún estadístico de orden. Se hace uso de las CDF´s de los positivos y negativos

$$F_P = \frac{1}{k}\sum\nolimits_{1\le i\le k} F_P^{(i)},\ F_N = \frac{1}{k}\sum\nolimits_{1\le i\le k} F_N^{(i)}$$

$F_P^{(i)}$, y $F_N^{(i)}$ son las CDF´s conjuntas para la medición de la i-ésima instancia. Este es robusto pero, si no se garantica la intercambiabilidad puede se no consistente.

En este contexto la prueba de la hipótesis nula

$H_0 : R_1 = \cdots = R_k$,

Se puede desarrollar la prueba usando el estadístico de prueba

$$S_h = \sum_{1 \le i \le k} d_n \left(\sqrt{n} \left(\hat{R}_i(t) - R_i(t) \right) \right)$$

En la fórmula intervienen una sucesión convergente de funciones reales $\{d_n\}_{n \in N}$ tal que $lim_{n \to \infty} d_n =_{cs} d$ y $R_i(t)$ es la curva ROC del i-ésimo sistema diagnóstico .

2.4. El ROC Paramétrico.

Es difícil conocer las verdaderas CDF`s envueltas en el estudio de desempeño de pruebas diagnósticas. Lo usual, al considerar métodos paramétricos, es que se acepte la normalidad de la función usada para evaluar el comportamiento de los tests diagnósticos. Así que se considera que esas funciones son normales, tanto para los decesos como para los no decesos, y que tienen diferentes medias.

Bajo tales supuestos los resultados son generados son continuos y

$$Y_1 \sim N(\mu_1, \sigma_1^2), Y_0 \sim N(\mu_0, \sigma_0^2)$$

Este es el marco usado generalmente, vea Ertugrul (2012), Molanes-López y Cao (2008).

Denótese Φ y ϕ como la CDF y la función probabilística de densidad de la normal respectivamente. Se espera que las medias sean diferentes.

Se toman muestras independientes $s_i, \|s_i\| = n_i, i = 1,0$, de los casos positivos (i=1) y negativos (i=0). Con ellas se estiman las medias y desviaciones estándar $(\hat{\mu}_i, \hat{\sigma}_i), i = 1,0$.

La curva ROC está dada por

$$ROC(t) = \phi(a + b\Phi^{-1}(t)), \qquad a = \frac{\mu_1 - \mu_0}{\sigma_1}, b = \frac{\sigma_0}{\sigma_1}$$

El área bajo la curva (AUC) es la probabilidad de que un sujeto positivo, seleccionado aleatoriamente, tenga un diagnóstico mayor que la de uno negativo y está dada por la función

$$AUC = \Phi\left(\frac{a}{\sqrt{1 + b^2}} \right)$$

La estimación de los parámetros se obtiene a partir de usar el método de Máxima Verosimilitud bajo la normalidad. Así que

$$\hat{a} = \frac{\bar{y}_1 - \bar{y}_0}{s_1}, \hat{b} = \frac{s_0}{s_1};$$

Donde

$$\bar{y}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_{ij}; s_j^2 = \frac{1}{n_j}\sum_{i=1}^{n_j}\left(y_{ij} - \bar{y}_j\right)^2 \; j = 0{,}1$$

Usando la regla de sustitución

$$\widehat{ROC}(t) = \phi\left(\hat{a} + \hat{b}\Phi^{-1}(t)\right)$$

La estimación de AUC por su parte es

$$\widehat{AUC} = \Phi\left(\frac{\hat{a}}{\sqrt{1+\hat{b}^2}}\right)$$

Las varianzas estimadas de los parámetros son derivadas usando el método Delta, siendo ellas

$$v(\hat{a}) = \frac{n_1(2+\hat{a}^2) + 2n_0\hat{b}^2}{2n_o n_1}, v(\hat{b}) = \frac{(n_1+n_0)\hat{b}^2}{2n_o n_1}$$

2.5. El ROC semi-paramétrico

La otra forma de tratar los problemas ROC es usar un enfoque semi-paramétrico. Estos métodos no hacen uso de hipótesis sobre las distribuciones de los resultados de los tests diagnósticos. Sin embargo, estiman los parámetros a y *b* como en el enfoque paramétrico.

Pepe (2000) aplicó Modelos Lineales Generalizados (GLIM) para ajustar curvas ROC. Estos permiten hacer inferencias usando Regresión Binaria. Estas técnicas son aplicadas a variables indicadoras construidas usando los resultados (Y_{i1} , Y_{j0}) provenientes de pruebas pareadas. Se toma la variable binaria

$$Z_{ij} = \begin{cases} 1 \; si \; Y_{i1} \geq Y_{j0} \\ 0 \; en \; otro \; caso \end{cases}; i = 1, \dots, n_0; i = 1, \dots, n_0$$

Los pares tienen el componente de un deceso y el de un no-deceso. Los parámetros de la curva ROC son obtenidos usando un método GLIM para la Regresión Binaria a partir de covariables. Así que se parte de todos los posibles pares de resultados diagnósticos $n_1 \times n_0$. Las tasas de falsos-positivos t_j son calculadas para estos pares usando los resultados de los no-decesos. Para cualquier par (Y_{i1} , Y_{j0}), se obtiene un $t_j \in \left\{\frac{1}{n_0}, \frac{2}{n_0}, \dots, 1\right\}$,

$$t_j = F_P(y_{0i})$$

La curva ROC se construye a partir de un método paramétrico al ajustar

$$g\left(ROC(t)\middle|\vec{B}\right) = \sum_{h=1}^{H} B_h g_h(t), \vec{B} = (B_1, \dots, B_H)$$

g es una función de conexión (link), g_h son ls funciones base y $\vec{B}$ un vector paramétrico desconocido. A partir del instrumental de los procedimientos GLIM se pude derivar

$$g\left(E(Z_{ij})\middle|\vec{B}\right) = \sum_{h=1}^{H} B_h g_h(t_i), \vec{B} = (B_1, \dots, B_H)$$

Al usar como función de linkeo los probits Φ^{-1} , $h_1(t_j) = 1$ and $h_2(t_j) = \Phi^{-1}(t_j)$, y el modelo es

$$E(Z_{ij}) = \Phi\left(B_1 + B_2\Phi^{-1}(t_j)\right)$$

Note que las estimaciones de los parámetros no dependen de hipótesis sobre los tests diagnósticos para hacer comparaciones. Sin embargo, los estimadores $\hat{B}_h, h = 1,2$, son obtenidos a partir de las herramientas de GLIM para la Regresión Binaria y a partir de ellas se estiman *a* y *b*.

En este contexto el AUC del modelo es

$$\widehat{AUC} = \Phi\left(\frac{\hat{B}_1}{\sqrt{1+\hat{B}_2^2}}\right)$$

Lo más usual es usar técnicas de re-muestreo para estimar las varianzas y hacer inferencias.

Gang et al. (1999) propusieron usar un enfoque a no-paramétrico para hacer las pruebas sobre las distribuciones de las poblaciones de decesos y de no-decesos para establecer cual es adecuada usando. Cai y Moskowitz (2004) por su parte propusieron estimar la curva ROC usando los Perfiles de Verosimilitud y los de Pseudo-Máxima Verosimilitud. Wan y Zhang (2007) hicieron uso de estimadores de las CDF´s basados en Kérneles.

3. ELEMENTOS COMPLEMENTARIOS DE ANÁLISIS.

3.1. Validación Cruzada y las curvas ROC.

La validación cruzada es un método simple para obtener adecuadas curvas ROC a partir de varias curvas. Usándolas se puede:

- Coleccionar las probabilidades para las instancias en las pruebas.
- Ordenar las instancias de acuerdo a las probabilidades.

Son de uso para desarrollar estrategias generales tomar conjuntos de instancias independientes y usarles para hacer una validación cruzada, hacer pruebas de hipótesis etc.

Haciendo uso de la validación cruzada se obtiene un sistema decisional más flexible. Vea la figura siguiente donde se ilustra como funcionaría tal procedimiento.

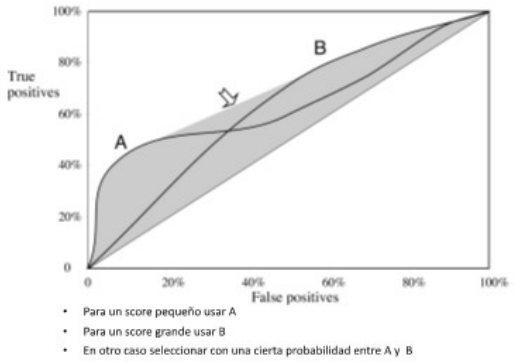


Ilustración 2:Curvas ROC para dos esquemas

Dados dos esquemas se obtendrá cualquier punto sobre el cono convexo generado al hacer la descripción siguiente:

- Usar las tasas TP y FP para el esquema 1: t_1 y f_1
- Usa las tasas TP y FP para el esquema 2: t_2 and f_2

Si el esquema 1 es usado en la predicción de N(1)=100×P % de los casos y para el esquema 2 es N(2)=100x(1-P)% , las tasas para la nueva curva determinada son

$TP = P\times t_1+(1-P) \times t_2$

$FP= P \times f_2+(1-P) \times f_2$

Notas sobre el uso de la validación cruzada

1- Este Evita la sobre ajuste.
2- Es muy recomendable para pequeñas datas.
3- No deben usarse pruebas para adecuar los parámetros.
4- Debe usarse una data separada e independiente para hace la validación.
5- Trate de introducir elementos de costo.

3.2. Clasificación sensible al costo.

En muchas aplicaciones es necesario considerar el costo asociado a una mala clasificación. Por ejemplo, no es lo mismo tener un error al clasificar un paciente como no afectado por cáncer que hacerlo con un usuario de internet.

Ejemplo. Se tienen dos matrices de confusión obtenidas bajo dos condiciones diferentes. Estos son:

Condición 1

	POSITIVO	NEGATIVO
POSITIVO	30	10
NEGATIVO	40	90

Condición 2

	POSITIVO	NEGATIVO
POSITIVO	10	10
NEGATIVO	20	110

Matriz de Costos

	POSITIVO	NEGATIVO
POSITIVO	0	2
NEGATIVO	5	0

Tasa de error de la Condición 1=50/170
Costo=10x2+40x5=220

Tasa de error de la Condición 2=30/150
Costo=10x2+20x5=120

Ilustración 3:Matrices de confusión

En casos donde la mala clasificacion tenga efectos sensibles, deben incorprarse elementos de costo cuando se hagan las predicciones. La idea básica es garantizar que la predicción de una clase con un "alto costo" solo se haga cuando se tenga mucha confianza en que sea correcta. Esta idea conlleva el tener conocimiento a apriorístico de las probabilidades de que una instancia pertenezca a cada clase. Es claro que a las clases más

verosímiles se asocien mayores probabilidades. Fijando c_i como el costo de la i-ésima clase y p_i como la probabilidad de que una instancia pertenezca a ella, el costo esperado es

$$\bar{c} = \sum_{i=1}^{I} c_i p_i$$

Se elegirá la clase que minimice el costo.

Los esquemas de aprendizaje generalmente buscan la minimización de la tasa de error total. Al hacer el entrenamiento los costos no serían tomados en cuenta. En esta fase los clasificadores generan sin considerar los costos de ser asignados a las clases, pero sirven para elecitar lo valores de las probabilidades. Herramientas usualmente usadas son los árboles de decisión, re-muestreo de instancias o su ponderación de acuerdo a los costos prefijados. Aunque en la práctica los costos no son conocidos el decisor puede hace una predicción de ellos.

Una visión adicional es obtenida al usar el mapeo de elevación (Lift Plot). Esta representa la derivativas de ploteo de la curva ROC. Observe en la figura siguiente un mapeo

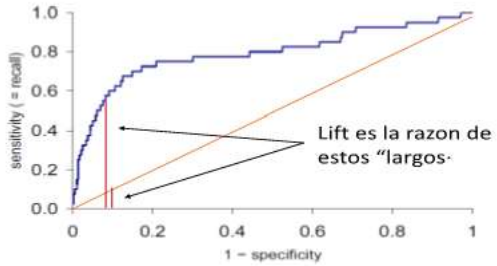


Ilustración 4: Ploteo de incrementos (Lift Plot)

Este comparar el desempeño del clasificador contra una ordenación aleatoria o contra otros clasificadores. Así que un mapeo de elevación permite visualizar y hacer una comparación de curvas ROC. Vea una comparación de 4 clasificadores en la figura siguiente.

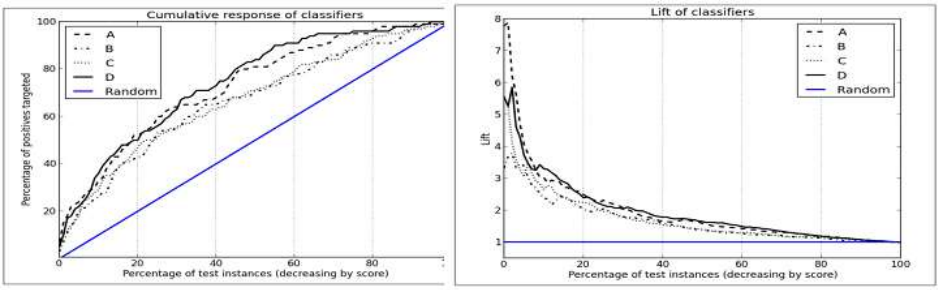


Ilustración 5: Comportamiento de los clasificadores

Este mapeo hace énfasis en la precisión inicial que es generalmente lo que solo interesa al decisor y en el desempeño en forma independiente. El factor de elevación da idea de cómo funciona la incrementación. Lo usual es que los puntos de ploteo se computen usando un espaciado regular, como 1/100, 1/1000 u otro. De no hacerse así y usar un valor inicial del mapeo demasiado grande habrá inestabilidad. Vea un ejemplo en la figura siguiente.

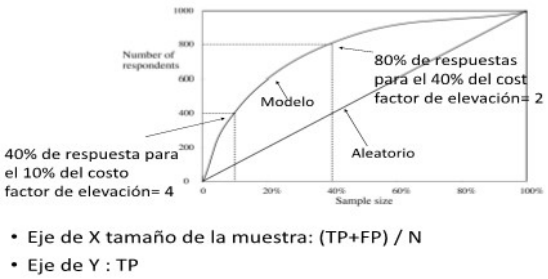


Ilustración 6:Un mapeo de la Elevación

3.3. Medidas usadas en el estudio de las curvas ROC

En algunas áreas al acudir al estudio de las curvas ROC son usadas medidas que son de interpretación propia de aspectos de la fenomenología particular. Algunas de ellas son

$$\text{Precisión} = \text{TP}/(\text{TP} + \text{FP})$$

Porcentaje de documentos relevantes retornados: $\text{recall} = \text{TP}/(\text{TP} + \text{FN}) = \text{TPR}$

o sea

Sensibilidad: $\text{TP} / (\text{TP} + \text{FN}) = \text{recall} = \text{TPR}$

$$\text{Medida} - \text{F} = (2\times\text{recall}\times\text{precisión})/(\text{recall} + \text{precisión})$$

Sumario de medidas: precisión promedio para $20\%, 50\%$ y 80% de recall

$$\text{Especificidad: TN} / (\text{FP} + \text{TN}) = 1 - \text{FPR}$$

$$\text{AUC} = \text{Área bajo la curva ROC}$$

Se pueden distinguir las temáticas en que se hace uso de esas medidas en áreas del conocimiento

Tabla 1: Temáticas por áreas de conocimiento

	Área de conocimiento	Denominación	Medida
Curva ROC	Comunicaciones	Tasas de TP y FP	$\frac{\text{TP}}{\text{TP} + \text{FN}}$ $\frac{\text{FP}}{\text{FP} + \text{TN}}$
Mapa de elevación	Mercadeo	TP tamaño de la muestra	$\frac{\text{TP} + \text{FP}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}$
Curva recall vs precisión	Recuperación de información	Recall Precisión	$\frac{\text{TP}}{\text{TP} + \text{FN}}$ $\frac{\text{FP}}{\text{TP} + \text{FP}}$

4. EJEMPLIFICANDO EL ANÁLISIS CON CURVAS ROC

4.1. Elementos de base

El nombre ROC proviene de la necesidad de interpretar las señales de radares durante la II Guerra Mundial. Era necesario identificar con exactitud si una señal identificaba aviones u otro objeto volador, pájaros, por ejemplo. A partir de las mediciones se obtendrá una función escalonada y esta, a veces se suaviza como se ejemplifica en la figura que se presenta debajo.

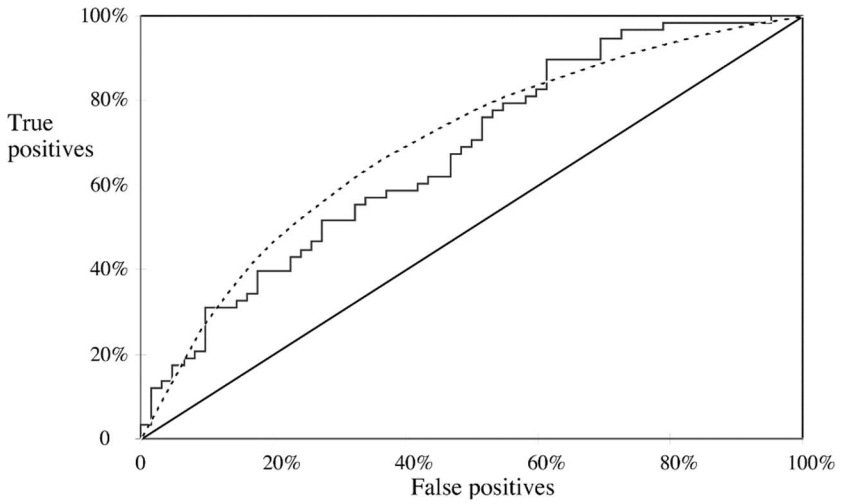


Ilustración 7:Cuerva de ROC suavizada

La necesidad de identificar señales pervade actualmente otras áreas no militares. Se ha tratado la problemática médica en biomedicina, evaluaciones de tratamientos alternativos, Biomarcadores e imageonología. Esta es quizás el área más conocida pero el mundo moderno plantea áreas donde se usan herramientas del ROC, como el Aprendizaje Automatizado (Machine Learning). Si se enfrenta la necesidad de decidir si la señal obtenida es positiva es importante. Por ejemplo, cuando se enfrenta analizar el desempeño de varios clasificadores. Este es igual al caso de los biomarcadores en medicina: se debe establecer no solo cuan buenos son sino cual es el mejor.

La experimentación parte de tomar una muestra de gran tamaño, las instancias observadas son clasificadas en positivas o negativas. El clasificador los clasificará. Lo ideal es que identifique perfectamente las instancias, pero esto es difícil. La muestra obtenida y clasifica correctamente y representa lo verdadero. Por tanto, un calificador es evaluado a partir de comparar su clasificación con lo real. En la tabla que presentamos aparece el reporte y el significado probabilístico de una prueba estadística.

Tabla 2: Tipos de errores en la clasificación

Muestra \ Clasificador	Acepta como positivo	Acepta como negativo
Identifica como positivo (D = 0)	☺ Especificidad (1-α)	X Error de tipo I (Falso) ($Nivel\ de\ significación\ \alpha$)
Identifica como negativo (D = 1)	Error tipo II (Falso) β	☺ Sensibilidad (potencia de la prueba $1-\beta$)

Al mirar esta tabla se tiene que, del experimento se obtiene una tabla de contingencia, denominada en este contexto a veces como Matriz de Confusión, de doble entrada como se ve en la próxima tabla al considerar que los resultados reales conforman una Regla de Oro o Gold Standard

Tabla 3: Matriz de Confusión

		Predicción Del Modelo	
		Clasificado como negativo	Clasificado como positivo
GOLD STANDARD (Lo Verdadero)	Clasificado como negativo	n_{AA} (verdaderos negativos)	n_{AB} (falsos positivos)
	Clasificado como positivo	n_{BA} (falsos negativos)	n_{BB} (verdaderos positivos)

La variable de interés tendrá, para cada clase, una distribución diferente en general. Vea una ilustración en la figura siguiente.

Ilustración 8: Un ejemplo del COVID 19

La decisión es tomada a partir de evaluar la verosimilitud o fuerza de la señal recibida. Esquemáticamente es lo que se representa en la figura siguiente.

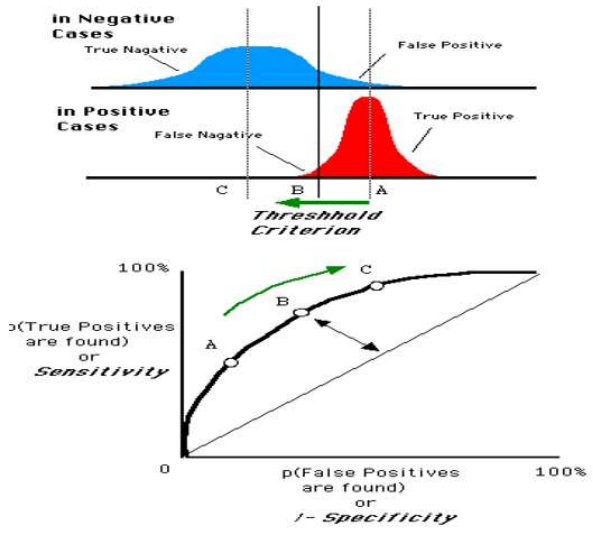


Ilustración 9: Distribución de la intensidad de la señal observada

Note que esto es similar a lo que soporta las pruebas de hipótesis en la estadística inferencial.

Es muy común que el especialista defina cual es el resultado de interés y que determinando un criterio de éxito/ fracaso. El resultado Y de la evaluación es generado por el análisis de un grupo de predictores $X_1,\ldots,X_k$. Un experimentó bien diseñado comenzará por determinar cuáles predictores son buenos para decidir si una instancia es un éxito o un fracaso. En la curva dada a continuación da una idea adicional sobre la geometría de una curva ROC.

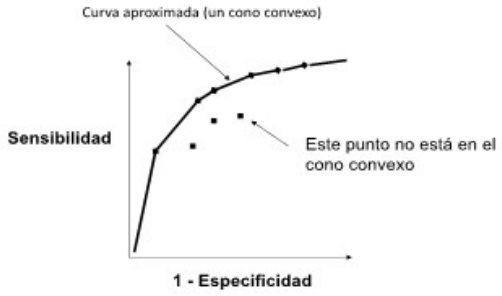


Ilustración 10: La geometría de una curva ROC

En el caso binario la variable de interés Y es dicotómica: si o no. Si Y es continua se determina uno o varios puntos de corte o umbrales. Vea un ejemplo en la figura siguiente, donde Y es el resultado del PCR en posibles infestados con el virus del COVID19. El umbral c determina si el individuo es considerado contagiado o no. Si Y>c es evaluado de positivo.

Es útil plantearse un grupo de umbrales $c_1, \ldots, c_k$. A partir de ellos se graduará la importancia del status de las instancias. Vea un ejemplo en el próximo gráfico.

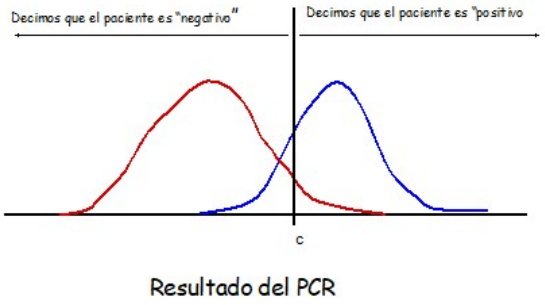


Ilustración 11: El Umbra

En el caso continuo se hará el análisis buscando un estimado de la curva ROC basándose en la experiencia empírica. Es posible hacer algún estudio usando modelos de regresión y hacer pruebas sobre el ajuste. Como anteriormente se vio también se puede modelar la curva ROC a través de covariables.

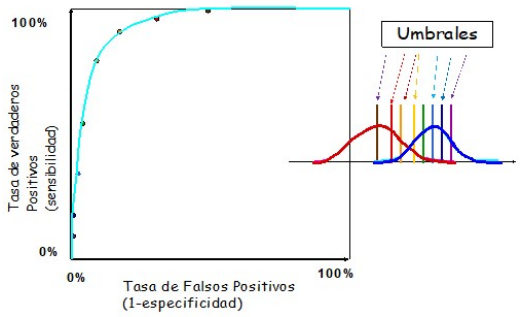


Ilustración 12: Curva ROC

El experimentador plotea los resultados observados creando una sucesión de puntos los que son unidos trazando una línea continua, como se ejemplifica continuación.

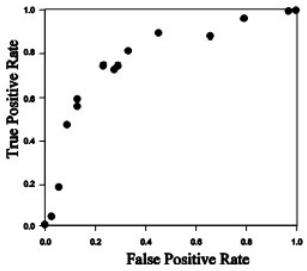


Ilustración 13: Una curva ROC

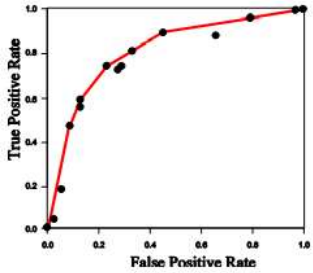


Ilustración 14: Observaciones

Vea en la figura siguiente el llamado espacio de definición de una curva ROC.

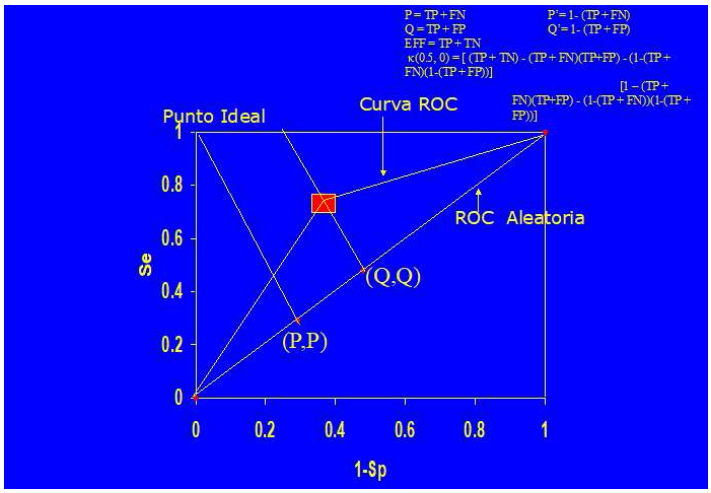


Ilustración 15: El espacio ROC

La valoración de una curva ROC es llevada a cabo analizando su gráfico en el espacio ROC. Vea una regla práctica para evaluar groseramente un test en la figura presentada a continuación.

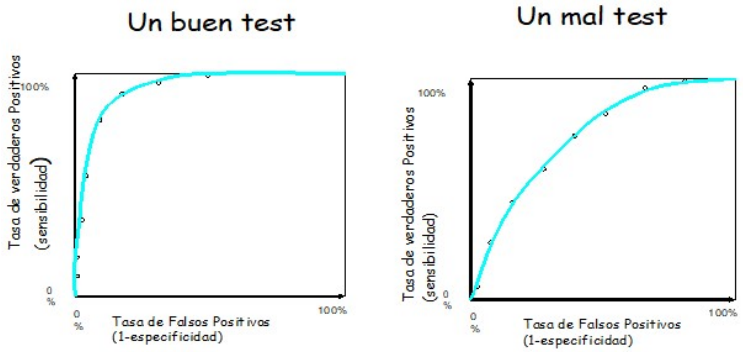


Ilustración 16: Evaluación de dos curvas ROC

4.2. Unas dudas estadísticas

La primera duda es si se pueden promediar unas curvas ROC. La respuesta es "sí". Vea la representación de la curva promedio de tres curvas ROC en la figura próxima. Esta es una salida de un computador usando datos obtenidos sobre el test de antígenos en personas sospechosas de ser portadoras del SARS-COV2.

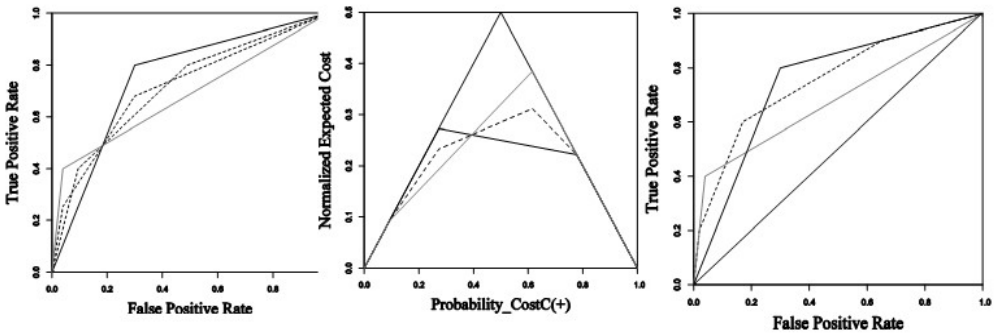


Ilustración 17: Promedio de curvas ROC

La segunda duda lleva a preguntar como computar un intervalo de confianza para el promedio de las curvas ROC. La solución es usar un método de re-muestreo y generar B matrices de confusión. Con ellas se puede computar una desviación estándar y dada la convergencia a la normal de los métodos de re-muestreo se computan las cotas superior e inferior del intervalo de confianza. Vea el ejemplo dado por la figura siguiente el gráfico de salida es la próxima figura, donde aparecen además algunas de las curvas ROC generadas por el re-muestreo

Verdadero	Predicción	
	pos	neg
positivo	78	22
negativo	40	60

TP = 0,78
FP = 0,4

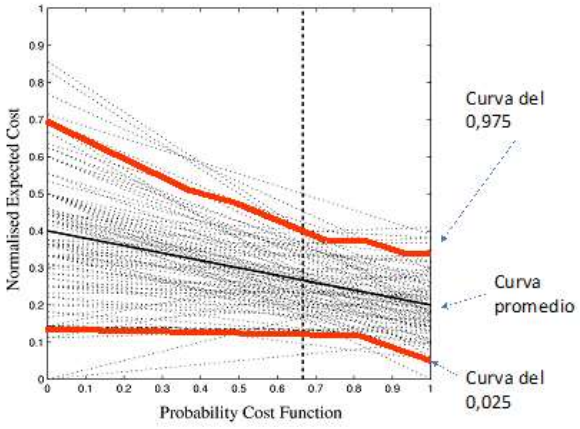


Ilustración 18: Ejemplos de intervalos de confianza del 0,95

La otra duda lleva a hacer la pregunta 3: ¿Como decir si dos curvas ROC difieren estadísticamente?

Para responder eso se deberá hacer un experimento pareado. Dos clasificadores hacen sus clasificaciones y puede aplicarse una prueba pareada. Las pruebas más adecuadas son del tipo no paramétrico, como se evidencia al consultar fuentes como Venkatraman (2000), Venkatraman y Begg (1996), Martínez-Camblor, et al. (2011), Bandos, et al. (2005) y Braun y Alonzo. (2008). Los datos de la tabla siguiente presentan las clasificaciones hechas por dos grafólogos, que identifican 100 documentos y les asignan a individuos.

Tabla 4: Clasificaciones

Predicciones del Clasificador 1	**Predicciones del Clasificador 2**	
	Pos.	**Neg.**
Positivo	**30**	**10**
Negativo	**0**	**60**

Para C1:FP1 = (30+10)/100 = 0,40

Para C2 FP2 = (30+0)/100 = 0,30

Y

FP2 – FP1 = -0,10

Se aplica un método de re-muestreo de esta matriz y se replica FP2-FP1.Se plotean los resultado para obtener un región de aceptación del 95% , como al desarrollar un intervalo de confianza. se obtiene un gráfico como el siguiente

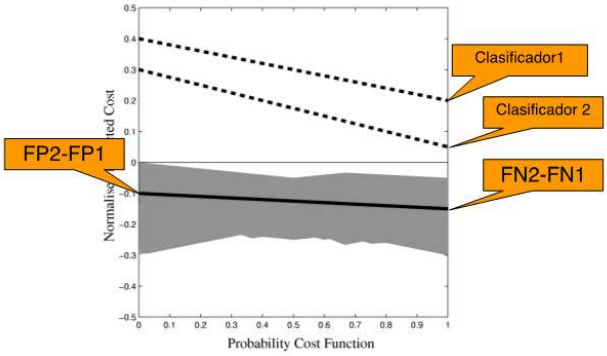


Ilustración 19: Remuestreo pareado para hacer prueba de significación

La significación se deriva de analizar el gráfico anterior. Vea en el siguiente gráfico la significación en forma gráfica,

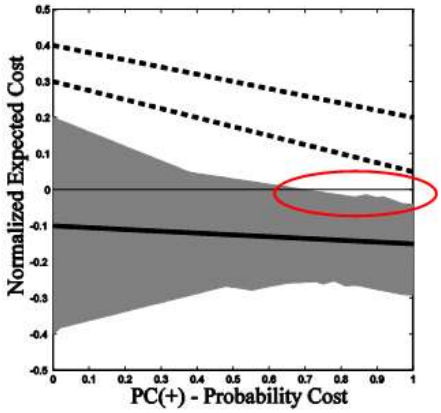


Ilustración 20: Un limitado rango de significación

Se puede decir que se darán las respuestas a estas preguntas haciendo los gráficos necesarios y hacer una inspección visual.

Una última pregunta será como establecer los errores asociados a procedimientos. Los errores pueden se valorados usando diversas medidas. Denótese que el valor real y/o la regla de oro genera la sucesión $v_1, \dots, v_n$ y las predicciones de ellas hechas por el procedimiento son $\hat{v}_1, \dots, \hat{v}_n$. Son usables, para evaluar la precisión del procedimiento de predicción, las medidas de error

$$\text{Error Cuadrático Medio} = \text{ECM} = \frac{1}{n}\sum_{j=1}^{n}(\hat{v}_j - v_j)^2$$

$$\text{Error Estándard} = \text{ES} = \sqrt{\frac{1}{n}\sum_{j=1}^{n}(\hat{v}_j - v_j)^2}$$

$$\text{Error Absoluto Medio} = \text{EAM} = \frac{1}{n}\sum_{j=1}^{n}|\hat{v}_j - v_j|$$

Es común querer evaluar la ganancia relativa de las predicciones respecto a la media

$$\text{Error Cuadrático Medio Relativo} = \text{ECMR} = \frac{\sum_{j=1}^{n}(\hat{v}_j - v_j)^2}{\sum_{j=1}^{n}(\bar{v}_j - v_j)^2}$$

$$\text{Error Estándard} = \text{ESR} = \sqrt{\text{ECMR}}$$

$$\text{Error Absoluto Medio} = \text{EAMR} = \frac{\sum_{j=1}^{n}|\hat{v}_j - v_j|}{\sum_{j=1}^{n}|\bar{v}_j - v_j|}$$

$$\text{Error Cuadrático Medio Relativo} = \text{ECMR} = \frac{\sum_{j=1}^{n}(\hat{v}_j - v_j)^2}{\sum_{j=1}^{n}(\bar{v}_j - v_j)^2}$$

El decisor debe establecer cual es la mejor medida o calcularlas todas para hacer una valoración integral.

Tabla 5: Tabla de medidas

	Actual	B1	B2	B3	B4
AUC	0,687	0,604	0,676	0,668	0,712
ES	11,5	17,8	21,7	43,3	17,4
EAM	10,6	11,3	28,5	43,4	19,2
ECMR	9,8%	12,2%	32,2%	39,4%	25,8%
EAMR	9,9%	13,1%	40,1%	24,8%	10,4%
CORRELACION	0,91	0,65	0,71	0,62	0,89

Ejemplo. Se evalúan 4 biomarcadores y el método actualmente en uso. Se mide la exactitud de ellos como predictores. Los resultados son los dados en la tabla anterior. En este caso el método actual es el mejor clasificador ingralmente y le sigue B4.

4.3. El AUC

El análisis de AUC tiene algunos problemas en sus interpretaciones, pues en general no tienen un significado claro en términos del problema específico. Una idea de cómo evaluar "cuantitativamente" el AUC es ejemplificado en la figura siguiente.

Una parte no despreciable del área proviene del rango de valores grandes de falsos positivos. Es necesario analizar regiones más restringidas por parte del experto. Como las curvas pueden cruzarse puede haber una diferencia significativa del comportamiento que no es explicada por las AUC.

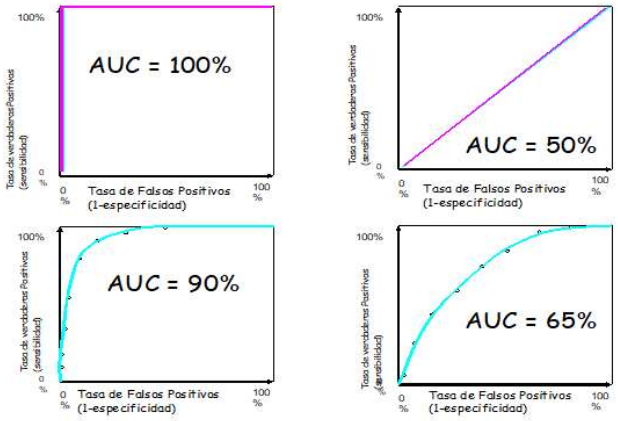


Ilustración 21: AUC para algunas curvas ROC

Se hace uso de uno de los tantos softwares donde se tienen implementados procedimientos para el análisis de las curvas ROC. Estos determinan sistemáticamente los mejores predictores y el índice de Youden que es:

Youden =(sensibilidad + especificidad) – 1.

Una regla práctica es considerar el patrón dado en la figura siguiente.

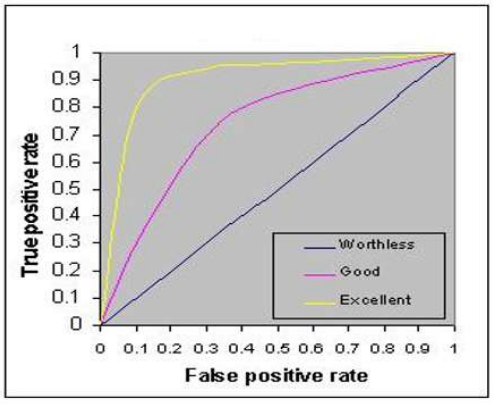


Ilustración 22. Comparando curvas ROC

También las salidas brindan resultados útiles para el análisis estadístico del AUC. Entre ellos los intervalos del coeficiente de 0,95 y algunos de ello el p-valor correspondiente para la prueba H0: AUC=0,5 vs H1:AUC>0,5. Si se rechaza el AUC es significativamente bueno. En general se brinda también la prueba Chi-cuadrado y el a p-valor se ofrece es también par la hipótesis nula AUC =0,5. Además el comportamiento de los ajustes se lleva a cabo usando comparaciones de la AUC estimadas a partir de hacer una prueba de un estadístico del tipo Mann-Whitney. Esta prueba es no-paramétrica y permite docimar la diferencia de la localización de dos poblaciones.

4.4. Ejemplo: Un problema Aprendizaje Automatizado

Se aborda la evaluación de clasificar de la información recabada para predecir el éxito de implantar un "extent" a un cardiópata. Se compara el desempeño de la regresión logística y las redes neuronales artificiales. Tradicionalmente los médicos recurren al uso de tablas de contingencia. Pero esto no permite hacer una reducción de los individuos que saldrán exitosamente de la intervención. Se trata con variables binarias y el interés es predecir el éxito y lo que se tiene son los conteos presentados en la tabla se de contingencia (de confusión). Estas son las tasas TP, FP, TN, FN. Las estadísticas a mano son:

FP+FN , TPR (Tasa de verdaderos Positivos): TP / (TP + FN) = Recall

FPR (Tasas de Falsos Positivos): FP / (TN + FP) = Precisión

Estos son usados en el análisis ROC para sumarizar y presentar el desempeño de un modelo de clasificación binario. Este modela las habilidades de distinguir entre los falsos y verdaderos positivos. En la table dada a continuación se brinda el resultado de la predicción de la probabilidad de éxito usando esto dos modelos. Valores cercanos a 1 se considerarían éxitos.

Tabla 6: Predicciones bajo los modelos

VERDADERO STATUS	PREDICCIÓN USANDO REGRESIÓN LOGISTICA	PREDICCIÓN USANDO REDES NEURONALES
1	0.7198	0.9038
0	0.2460	0.8455
0	0.1219	0.4655
0	0.1560	0.3204
0	0.7527	0.2491
1	0.3064	0.7129
0	0.7194	0.4983
0	0.5531	0.6513
1	0.2173	0.3806
0	0.0839	0.1619
1	0.8429	0.7028

Se fijaron varios umbrales y se obtuvo la próxima tabla

Tabla 7: Resultados con umbrales fijo

	PREDICCIÓN USANDO REGRESIÓN LOGISTICA		PREDICCIÓN USANDO REDES NEURONALES	
Umbral	Tasas de verdaderos (TP)	Tasas de falsos (FP)	Tasas de verdaderos (TP)	Tasas de falsos (FP)
1	1	1	1	1
0,9	1	0,8571	1	1
0,8	1	0,5714	1	0,8571
0,7	0,75	0,4286	1	0,7143
0,6	0,5	0,4286	0,75	0,5714
0,5	0,5	0,4286	0,75	0,2857
0,4	0,5	0,2857	0,75	0,2857
0,3	0,5	0,2857	0,75	0,1429
0,2	0,25	0	0,25	0,1429
0,1	0	0	0.25	0
0	0	0	0	0

Usando las herramientas ROC se obtiene la representación gráfica siguiente:

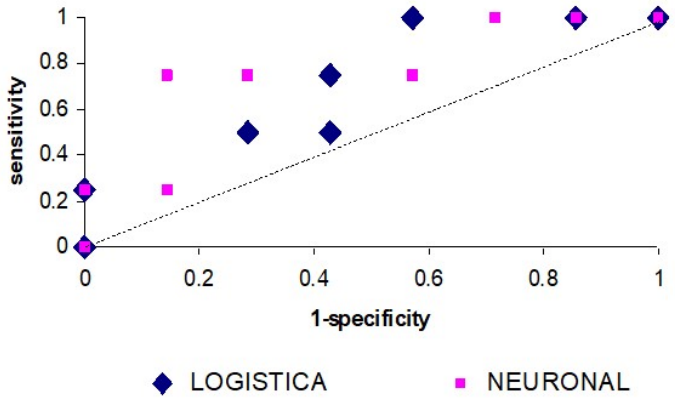


Ilustración 23: Curva ROC

Para calcular las AUC`s y usando la regla de la cuadratura trapezoidal para hacer el cálculo del área se obtiene la tabla dada a continuación.

- AUC
  - Usando la regla trapezoidal para calcular el area
- *Conclusión: mejor la Red Neuronal Network.*

	AUC
LOGISTICA	0.7321
NEURONAL	*0.7679*

Ilustración 24: Cuantificación ROC

Los resultados son muy similares pero las redes neuronales son mejores.

El análisis desarrollado sigue las normas sugeridas en Albert et al. (2014) y Pepe (2000, 2003).

4.5. Un ejemplo: Clasificación de tumores por 3 médicos

Se enfrentan 3 médicos a la evaluación de imágenes obtenidas de tumores. Ellos deben clasificarles como benignos o cancerígenos. Los pacientes son operados y se da la evaluación de su real condición. Vea los resultados a continuación.

verdadero	Predicción pos	Predicción neg
pos	40	60
neg	30	70

Médico 1
TP = 0,4
FP = 0,3

verdadero	Predicción pos	Predicción neg
pos	70	30
neg	50	50

Médico 2
TP = 0,7
FP = 0,5

verdadero	Predicción pos	Predicción neg
pos	60	40
neg	20	80

Médico 3
TP = 0,6
FP = 0,2

Ilustración 25: Médicos clasifican tumores en cancerígenos o benignos

Una mirada a la modelación en forma gráfica que servirá de patrón para hacer la evaluación es la esquematizada en la figura dada a continuación.

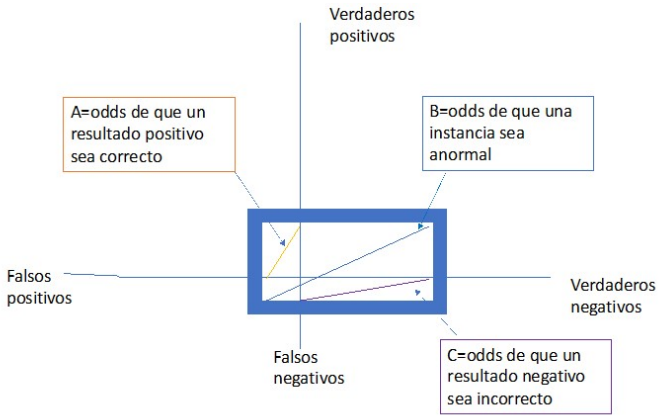


Ilustración 26: Trayectoria del test y la aleatoria

Un test "funciona" si es mejor que si se decidiera aleatoriamente. Mientras más alejado en el plano superior mejor. La representación de los resultados en el plano es dada en la figura siguiente.

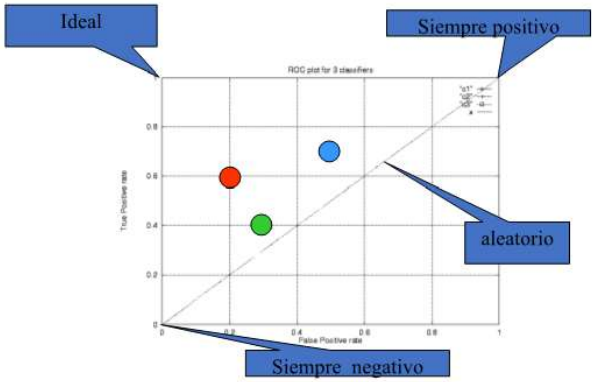


Ilustración 27: Curva ROC de los tres médicos

Como se ve no se puede decir que las clasificaciones se han hecho en forma aleatoria. El comportamiento es mejor. Estas evaluaciones pueden ser analizadas geométricamente.

Vea en la próxima figura el rango donde se mueven las clasificaciones. Así que se puede aceptar que los clasificadores son mejores que hacerlo aleatoriamente.

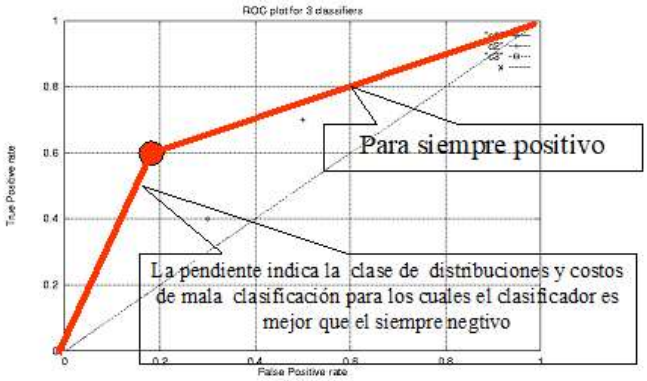


Ilustración 28: Rango de operación

La figura siguiente establece esta para los médicos 2 y 3. El decisor puede, al analizarla, establecer porque el médico 3 es el mejor clasificando.

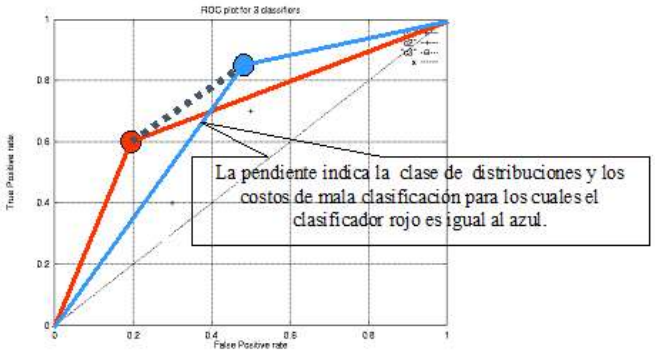


Ilustración 29: Cono convexo

Podemos establecer además la dominancia de los 3 clasificadores y representarles en la figura siguiente en la que es ve que el medido 3 "domina" a los demás. O sea que son más confiables sus clasificaciones

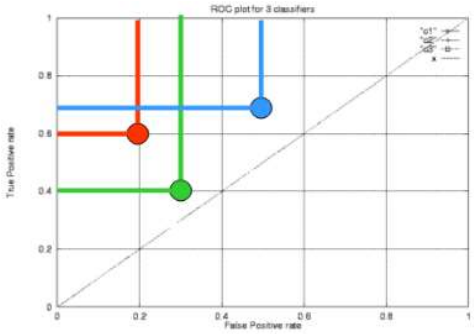


Ilustración 30: Dominancia

Así que se concluirá que el análisis de las imágenes provee de una adecuada valoración de la naturaleza del tumor. Entre los médicos se prefirió el número tres.

4.6. Un Ejemplo: Una visita usando costos

Tomemos en consideración un modelo estándar, en el que se fija el costo de una correcta clasificación: C_0=0. El costo de una mala clasificación dependerá solo de la clase y no del individuo clasificado. Tomaremos que los costos son aditivos. Se debe fijar que los costos de clasificación no son conocidos exactamente al hacer la evaluación pudiendo variar con el tiempo y del clasificador. Por su parte los falsos positivos (FP) y verdaderos positivos (TP) no varían con el tiempo y no la ubicación pudiendo ser estimados con exactitud.

En este contexto se pueden usar medidas escalares para evaluar el desempeño de los clasificadores. Estas son fundamentalmente:

- Costo Esperado
- Curvas ROC
- Área bajo la curva ROC
- Exactitud
- Técnicas de Visualización

Sin embargo, estas no brindan toda la información suficiente para evaluar los resultados observados de las clasificaciones. Los valores que caracterizan estas son FP y TP. Quedan dudas sobre como se distribuyen los errores entre las clases, se desempeñan los clasificadores ante distintas condiciones. Las medidas escalares requieren de establecer alguna ordenación de los clasificadores. Estos son tabulables pero no son sino un conjunto de valores que pueden no tener un impacto en transmitir la realidad al presentar un paper o dar una charla.

Por ello el uso de métodos visuales, como la forma de las curvas, pueden ser más informativos y complementar la información brindada por los escalares. El espacio es dado en la figura siguiente.

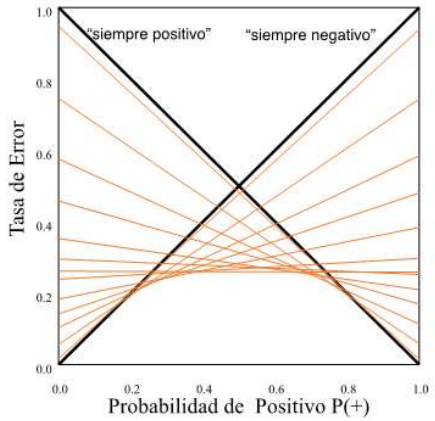


Ilustración 31. Una mirada a las Curvas de Costo

Al incluir costos se tienen

- Costos posibles por la mala clasificación bajo ciertas condiciones.
- Las posibles tasas posibles bajo ciertas condiciones
- Las condiciones bajo las cuales una curva se desempeña mejor que otra.

Vea la figura próxima en la que se analizan los costos de dos curvas

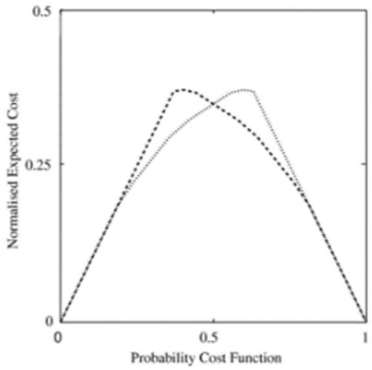


Ilustración 32: Comparando curvas de costo

Al tomar en cuenta las curvas de costo se incorpora la variable

$$Y = FN \times \vartheta + FP(1-\vartheta);\ \vartheta = \frac{P(+)C(-|+)}{P(+)C(-|+) + (1-P(+))C(+|-)}$$

Esta es el costo esperado normalizado a [0,1].

Veamos el análisis usando el ejemplo desarrollado 3.5.

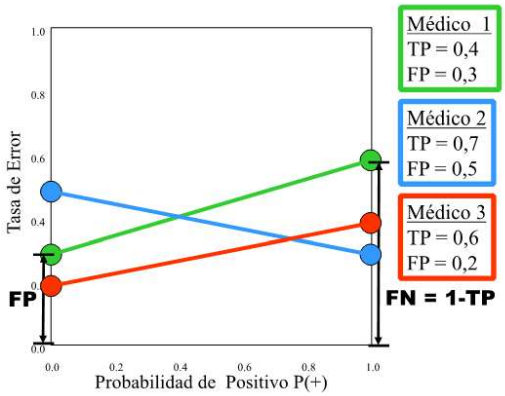


Ilustración 33: Curvas de Costo para los médicos clasificando tumores

Vea la comparación de los médicos 2 y 3 en la figura siguiente.

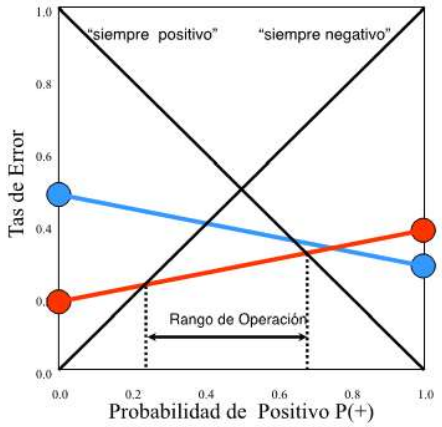


Ilustración 34: Analizando las curvas de costos de los médicos clasificando tumores

Una mirada geométrica se brinda en la próxima figura

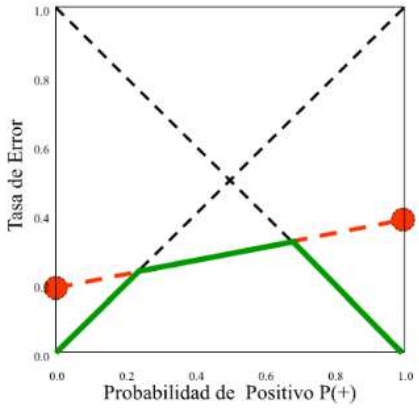


Ilustración 35: Envoltura inferior

Vea como el estudio desarrollado en 3.5 es complementado con este.

4.7. Un Ejemplo: identificación de las secuelas de recuperados de COVID-19

El uso de marcadores es de interés para establecer si un paciente al estar recuperado tiene necesidad de un tratamiento especializado para recuperar la funcionalidad de sus pulmones. Se desea establecer que marcador usar para hacer la diagnosis. Un problema fundamental es establecer con exactitud al punto de corte optimal. Acudiendo a las curvas ROC ayudará a establecer las exactitudes de los diagnósticos que se obtienen de los marcadores. Para ello se deben comparar varias curvas alternativas y seleccionar el punto de corte. Siguiendo la metodología usada corrientemente, como se puede ver en Malin (2005) y Alonzo y Pepe (2002), se requiere de tener una matriz de confusión como la de la tabla siguiente.

Tabla 8: Matriz de confusión

Resultado de usar el punto de corte	Resultado de la estándar	regla de oro
	positivo	negativo
Positivo	TP (verdadero positivo)	FP (falso positivo)
negativo	FN (falso negativo)	TN (verdadero negativo)

Y como se sabe

$$S = Sensibilidad\ = \frac{TP}{TP + FN}$$
$$E = Especificidad\ = \frac{TN}{FP + TN}$$
$$IY = Indice\ de\ Youden = S + E - 1$$

La curva ROC es determinada al plotear FP vs TP. Su exactitud es medida mediante el AUC y el punto de corte (c) es evaluado a partir de IY.

El estudio se basa en 750 pacientes adultos de ambos sexos. El índice de peligro de complicaciones pulmonares (IP) es explorado midiéndose la funcionalidad respiratoria (X) y la percepción de molestias por el sujeto (Q). Se indagará sobre la potencialidad de usar X y Q para cada paciente. IP es la medida de predicción

$$IP(j) = \begin{cases} 1\ si\ el\ indice\ de\ peligrosidad\ es\ superado \\ 0\ en\ otro\ caso \end{cases}; \forall j: E\big(IP(j)\big) = \pi$$

Así que cada paciente (instancia) se asocia al vector $\big(IP_j, X_j, Q_j\big)$. Usando procedimientos del SAS

PROC LOGISTIC para la comparación de curvas se declaran

$$X, Q\ como\ variables\ independientes$$

Veamos los análisis realizados a partir de considerar el modelo logístico

$$ecuación\ logística: ln\left(\frac{\pi}{1-\pi}\right) = B_0 + B_1 X + B_2 Q$$

Tabla 9: Análisis de los estimadores máximos verosímiles

Parámetro	GL	estimados	Error estándar	p-valor	Wald Chi-cuadrado	IC(0,95)
Intercepto	1	-2,5918	0,3034	<0,001	38,831	-2,5988 -1,9852
X	1	0,4141	0,2070	0,028	10,264	0,0010 0 ,8255
Q	1	0,6543	0,2833	<0,001	27,363	0,0884 1,3216

O sea que todas las variables son significativas para hacer la diagnosis.

Se tiene que

Tabla 10: Indicadores para evaluar el modelo

Youden index

Youden index J	0.8202
95% Confidence interval[a]	0.6979 to 0.9051
Associated criterion	>108.9
95% Confidence interval[a]	>107.9 to >114.8
Sensitivity	90.91
Specificity	91.11

[a] BC_a bootstrap confidence interval (1000 iterations; random number seed: 978).

Los médicos se cuestionaron si no es suficiente considerar la funcionalidad. En tal caso

Tabla 11: Análisis de los estimadores máximos verosímiles

Parámetro	GL	estimados	Error estándar	p-valor	Wald Chi-cuadrado	IC(0,95)
Intercepto	1	-2,3668	0,1804	<0,0001	30,73	-2,7276 -2,0060
X	1	0,6501	0,0147	0,0017	9,87	0,6207 0,6795

O sea que sí, es aceptable usar solo X para hacer las evaluaciones.

Otro análisis es el de las relaciones que se presentan a continuación.

Tabla 12: Análisis de los estadísticos ROC de asociación

Modelos ROC	AUC	Error estándar	IC del 95% de Wald	D de Gini	Gamma	Tau-a
Modelo	0,884	0,0241	0,828-0,940	0,777	0,777	0,364
X	0,732	0,0421	0,741-0,892	0,453	0,453	0,246
Q	0,883	0,0372	0,792-0,919	0,738	0,738	0,377

Como las AUC estimadas generan regiones de aceptación (IC`s) no se acepta H0: AUC=0,5 y todas son mejores que usar un clasificador aleatorio.

Para comparar las curvas ROC se analiza la salida de los contrastes. Vea la tabla siguiente.

Tabla 13: Contrastes ROC de las estimaciones y tests

Contraste	Estimación	Error estándar	IC del 95% de Wald	Chi-cuadrado	p-valor
Modelo-X	0,162	0,041	0,082-0,241	25,976	0,000
Q-X	0,134	0,053	0,030-0,238	6,374	0.0111

En resumen, al comparar Q con X, como p=0,0111, el AUC de Q, que como se vio es 0,883, es significativamente mayor que la de X, que es 0,732.

Ahora

Tabla 14: Análisis de los estadísticos ROC de asociación

Modelo ROC	AUC	Error estándar	IC del 95% de Wald	D de Gini	Gamma	Tau-a
Modelo	0,8841	0,0319	0,783-0,918	0,777	0,777	0,388
IP	0,00	0,000	0,500-0,500	0,002	0,002	0,000

Así que la curva ROC es estadísticamente significativa.

Tabla 15: Resultados de la prueba de contraste ROC

	df	Chi-cuadrado	p-valor
Referencia=modelo	1	220,562	<0,0001

Por otra parte, si usamos solo Q

Tabla 16: Análisis de los estimadores máximos verosímiles de Q

Parámetro	GL	estimados	Error estándar	Wald Chi-cuadrado	p-valor Chi-cuadrado
Intercepto	1	-2,6530	0,4198	39,707	<0,0001
Q	1	0,1878	0,0317	31,196	<0,0001

El nuevo modelo es también adecuado para hacer la predicción y es

$$logit = -2{,}6530 + 0{,}1878\text{Q}$$

Analizando los niveles de probabilidad para predicción de las instancias. La variable de contraste (umbral) es obtenida al resolver la regresión logística para resolver para Q=c. Los logits para el correspondiente c y IY son calculados y ordenados ascendentemente en términos de IY (índice de Youden). Valores altos de IY sugieren mejores umbrales para el score. En este caso se puede aceptar c=10,384 pues para ese se obtuvo el máximo valor de IY. El decisor utilizará entonces con regla de clasificación "si Q>10,38 la instancia se considera positiva".

References

Albert, P., S., Liu and Aiyi, N. T. (2014): Efficient logistic regression designs under an imperfect population identifier. Biometrics. 70,175–184.

Alonzo T. A. and M. S. Pepe (2002): Distribution-free ROC analysis using binary regression techniques. Biostatistics,. 3, 421–432.

Bandos, A. I., Rockette, H. E. and Gur, D. (2005): A permutation test sensitive to differences in areas for comparing ROC curves from paired design. Statistics in Medicine, 24, 2873–2893.

Braun, T. and Alonzo, T. A. (2008): A modified sign test for comparing paired ROC curves. Biostatistics, 9, 364–372.

Cai T. and C. S. Moskowitz (2004): Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. Biostatistics, 5, 573–586.

Carleos, C., Martínez-Camblor, P. and Corral, N. (2010): ROCtest: R package for flexible ROC curve comparison. In ERCIM'10. London, UK

Cerda J. y Cifuentes L. (2012):Uso de curvas ROC en investigación clínica. Aspectos teóricos-prácticos. Rev Chil Infect; 29, 138-141.

Dorfman, D.D. (2006). ROC Software Listing. [Software Website] http://perception.radiology.uiowa.edu/Softare/ReceiverOperatingCharacteristicROC/tabid/120/Default.aspx

Drummond C and Holte R. (2004): What ROC curves can and can't do (and cost curves can). In Proceedings of the Workshop on ROC Analysis in AI; in conjunction with the European Conference on AI. Valencia, Spain. 2004.

Drummond C. and R. C. Holte. (2006): Cost curves: An improved method for visualizing classifier performance. Machine Learning, 65, 95–130.

Ertugrul C., F. Mutlu, C. Bal, S. Oner, K. Ozdamar, B. Gok and Y. Cavusoglu (2012):

Comparison of Semiparametric, Parametric, and Nonparametric ROC Analysis for Continuous Diagnostic Tests Using a Simulation Study and Acute Coronary Syndrome Data. Computational and Mathematical Methods in Medicine, Article ID 698320.

Fawcett, T. (2003): ROC Graphs: Notes and Practical Considerations for Data Mining Researchers. Intelligent Enterprise Technologies Laboratory HP Laboratories Palo Alto HPL-2003-4

Fawcett, T. (2006): An introduction to ROC analysis. Pattern Recognition Letters 27, 861–874.

Gang, L. I. , R. C. Tiwari, and M. T. Wells (1999): Semiparametric inference for a quantile comparison function with applications to receiver operating characteristic curves. Biometrika, 86,487–502.

Hui, S. L., Zhou, X. (1998): Evaluation of diagnostic tests without gold standards. Statistical Methods in Medical Research. 7:354–370.

Hsieh, F., & Turnbull, B. W. (1996). Nonparametric and semiparametric estimation of the receiver operating characteristic curve. Annals of Statistics, 24, 25–50.

Krzanowski, W. J. and D. J. Hand (2009): ROC Curves for Continuous Data. Chapman & Hall/CRC Monographs on Statistics and Applied Probability.

Lloyd, C. J. (1998). Using smoothed Receiver Operating Characteristic curves to summarize and compare diagnostic systems. Journal of the American Statistical

Association, 93, 1356–1364.

Lloyd C. J. and Z. Yong (1999): Kernel estimators of the ROC curve are better than empirical. Statistics and Probability Letters, 44, 221–228.

Macaskill P, Gatsonis C, Deeks J. J, Harbord R. M. and Y. Takwoingi (2010): Analysing and Presenting Results. In: Deeks JJ, Bossuyt PM, Gatsonis C (editors), Cochrane Handbook for Systematic Reviews of Diagnostic Test Accuracy Version 1.0. The Cochrane Collaboration, 2010. Available from: http://srdta.cochrane.org/.

Malin B. (2005): Probabilistic prediction of myocardial infarction: logistic regression versus simple neural networks. *Data Privacy Lab Working Paper WP-25, School of Computer Science, Carnegie Mellon University*.

Martínez-Camblor, P. (2007). Comparación de pruebas diagnósticas desde la curva ROC. Revista Colombiana de Estadística, 30(2), 163–176.

Martínez-Camblor, P., Carleos, C., & Corral, N. (2011a). Powerful nonparametric statistics to compare k independent ROC curves. Journal of Applied Statistics, 38, 1317–1332.

Metz, C. E. (1978): Basic principles of ROC analysis, Seminars in Nuclear Medicine, 8, 283–298.

Metz, C.E. (2011). Metz ROC Software at the University of Chicago [Software Website] http://metz-roc.uchicago.edu

Molanes-López, E. M., y Cao, R. (2008). Relative density estimation for left truncated and right censored data. Journal of Nonparametric Statistics, 20, 693–720.

Pardo M.C. , A.M. Franco-Pereira (2017): Non Parametric Roc Summary Statistics. Revstat – Statistical Journal 15, 583–600.

Peizhou L., H. Wu, and T. Yu (2017): ROC Curve Analysis in the Presence of Imperfect Reference Standards. Stat Biosci. 9, 91–104. .

Pepe, M. S. (2000): An interpretation for the ROC curve and inference using GLM procedures. Biometrics, 56,. 352–359, 2000.

Pepe, M. S. (2003): The Statistical Evaluation of Medical Tests for Classification and Prediction, Oxford University Press, New York.

Qiu, P., & Le, C. (2001). ROC curve estimation based on local smoothing. Journal of Statistical Computation and Simulation, 70(1), 55–69.

Venkatraman, E. S. (2000). A permutation test to compare receiver operating characteristic curves. Biometrics, 56, 1134–1138.

Venkatraman, E. S., & Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment.

Biometrika, 83(4), 835–848.

Wan S. and B. Zhang (2007): Smooth semiparametric receiver operating characteristic curves for continuous diagnostic tests. Statistics in Medicine, 26, 2565–2586.

Zou, K. H., W. J. Hall, and D. E. Shapiro (1997): Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. Statistics in Medicine, 16, 2143–2156.

Capítulo 6 pp 87-96

MÉTODO PARA DETERMINAR LAS FASES MINERALES DEL CLÍNKER Y SU INFLUENCIA EN REDUCIR LOS DAÑOS AL MEDIO AMBIENTE

METHOD TO DETERMINE THE MINERAL PHASES OF THE CLINKER AND HIS INFLUENCE IN REDUCING THE DAMAGES TO THE ENVIRONMENT

Carlos Alberto Álvarez Bravo[1], Manuel E. Cortés Cortés[1] y Mario Moreira[2]
[1]Dpto. de Matemática, Facultad de Ingeniería, Universidad de Cienfuegos, Cienfuegos, Cuba
[2]Ingeniero de Proceso y Mejora Continua, Cementos Cienfuegos S.A, Cienfuegos, Cuba,

RESUMEN

El cálculo de los porcentajes de las fases minerales del clínker en la producción de cemento es un tema interesante y de gran utilidad en los momentos actuales, el objetivo fundamental de la siguiente investigación es reformular las Ecuaciones de Bogue para calcular el porcentaje de las fases minerales del clínker en la producción de cemento de la Empresa Mixta "Cementos Cienfuegos S.A" (CCSA) ubicada en la provincia de Cienfuegos, y con ello influir en la reducción de los daños al Medio Ambiente derivados del proceso de fabricación de dicho producto. El cálculo se realiza a partir del conocimiento de la composición de los principales óxidos (CaO, SiO_2, Al_2O_3, Fe_2O_3 y SO_3) en el clínker, pero el interés de los especialistas es efectuar el mismo antes de la obtención del clínker partiendo de los porcentajes de los óxidos en la materia prima para tomar las decisiones en el proceso de producción del mismo, garantizando la calidad del producto y una disminución en las pérdidas por producciones de baja calidad y en los daños al Medio Ambiente, de ahí la necesidad de estimar los modelos de la relación entre los porcentajes de cada óxido en la materia prima y en el clínker. En base a estas relaciones se modifican las ecuaciones de Bogue para obtener los porcentajes de las fases minerales del clínker en función de los porcentajes de los principales óxidos en la materia prima, antes de la producción del mismo. Los modelos estimados, de gran utilidad en el control de la calidad del clínker se corroboran en cuanto a su validez estadística, estabilidad y precisión, influyendo en la reducción de los daños al Medio Ambiente derivados del proceso de fabricación de dicho producto.

ABSTRACT

The calculation of the percentages of the mineral phases of the clínker in the production of cement is an interesting and great- benefit theme in the present-day moments, the fundamental objective of the following investigation is rephrasing Bogue's Equations to calculate the percentage of the mineral phases of the clinker in the production of cement of the Mixed Enterprise Cementos Cienfuegos S.A (CCSA) located in the province of Cienfuegos, and with it influencing the reduction of the damages to the Environment derived of the manufacturing process of the aforementioned product. Calculation comes true from the knowledge of the composition of the main oxides (CaO, SiO2, Al2O3, Fe2O3 y SO3) in the clínker, but the interest of the specialists is realizing the same before the obtaining of the clínker and from the percentages of the oxides in the raw material to take the decisions in the process of production of the same, guaranteeing the quality of the product and a decrease in the losses for low-quality productions and in the damages to the Environment, from there the need to value the models of the relation between the percentages of each Oxide in the raw material and in the clínker. On the basis of these relations Bogue's equations to obtain the percentages of the mineral phases of the clínker in terms of the percentages of the main oxides in the raw material get modified, before the production of the same. The models valued, of great benefit in the control of the quality of the clínker

[1] calvarez@ucf.edu.cu, mcortes@ucf.edu.cu, mmoreira@cementoscfg.com

corroborate themselves as to their statistical validity, stability and precision, influencing the reduction of the damages to the Environment derived of the manufacturing process of the afore.

PALBARAS CLAVES: Clínker, **c**álculo de las fases minerales del clínker, Ecuaciones de Bogue, Regresión lineal.

KEY WORDS: Clinker, Calculation of clinker mineral phases, Bogue's equations, Linear Regression.

INTRODUCCIÓN

Según [1] ¨el cemento Portland es un producto finamente molido, altamente reactivo y compuesto de clínker, yeso y algunos materiales de adición¨ Para la producción de cemento se necesita que ocurran una serie de procesos físico químicos que dan origen a un producto intermedio extraordinariamente valioso llamado clínker que representa las combinaciones estequiométricas de los óxidos fundamentales de las materias primas, a este proceso se le denomina proceso de clinkerización, proceso que solo tiene lugar en determinadas condiciones que están creadas previamente en el horno a temperaturas cercanas a los 1450 C^{o}.

Generalmente, las materias primas (piedra calcárea y materiales arcillosos) para la producción de cemento proceden de recursos no renovables y su extracción tiene un notable impacto ambiental, como suele suceder con todas las extracciones de minerales. En lo referente al proceso industrial, la obtención del clínker es la parte más importante en la fabricación de cemento e implica un elevado consumo de energía y, posteriormente, emisiones importantes de gases y polvo al molerlo [2].

La eficiencia del proceso de clinkerización, (o calidad del clínker a la salida del horno), depende de la temperatura del horno, el tiempo de residencia de la harina (materias primas homogeneizadas) en el horno y los parámetros de calidad de la harina, o sea, la granulometría y la correcta composición de los óxidos: CaO; Fe_2O_3; Si_2O, Al_2O_3y SO_3 [1].

Las desviaciones en los parámetros de calidad de la harina que entra al horno (considerando la granulometría constante) implican variaciones en el proceso de cocción y por tanto variaciones de la eficiencia del proceso de clinkerización en cuanto a lo que se refiere a las principales “fases minerales” constituyentes esenciales del clínker. Las principales fases en el clínker son: alita (C3S), belita (C2S), celita (C3A) y ferrita (C4AF), además, pueden estar presentes cristales de cal libre, periclasa y sulfatos alcalinos, entre otros [1]. Las proporciones, la cristalinidad y la textura de estas fases minerales en el clínker controlan propiedades tan importantes en el cemento como: fraguado, calor de hidratación, reactividad y desarrollo de resistencias [3, 4]. De ahí la importancia de cuantificarlas con precisión para así evitar que el producto tenga que ser producido reiteradamente y con ello impedir la extracción de recursos no renovables, un elevado consumo de energía y emisiones de gases contaminantes nuevamente, todo lo cual es perjudicial al Medio Ambiente.

R. H. Bogue desarrolló un proceso de cálculo según el cual, a partir del análisis químico, se puede calcular el contenido en minerales del clínker (en porcentaje), sobre todo, de alita (*C*3*S*), belita (*C*2*S*), celita (*C*3*A*) y ferrita (*C*4*AF*). A las ecuaciones encontradas por Bogue se les conoce en la actualidad como las “Ecuaciones de Bogue” [5]. Estas ecuaciones fueron planteadas así:

$$C3S = 4.071C_aO - \left(7.6S_iO_2 + 6.718Al_2O_3 + 1.43Fe_2O_3 + 2.852SO_3\right) (1)$$

$$C2S = 2.867S_iO_2 - 0.7544C_3S \ (2)$$

$$C3A = 2.65Al_2O_3 - 1.692Fe_2O_3 \ (3)$$

$$C4AF = 3.043Fe_2O_3 \ (4)$$

El cálculo del contenido en minerales del clínker mediante las Ecuaciones de Bogue se realiza a partir del conocimiento de la composición de los principales óxidos (CaO, SiO_2, Al_2O_3, Fe_2O_3 y SO_3) en el clínker, por tanto si se tiene en cuenta que el horno instalado es de 220 ton/h de clínker (con una significativa inercia) es de interés para los especialistas de Cementos Cienfuegos S.A realizar el cálculo antes de la obtención del clínker y de ahí la necesidad de predecir los porcentajes de los óxidos en el clínker, lo cual se realiza a partir de encontrar como depende el porcentaje de cada uno de los óxidos en el clínker del porcentaje de estos óxidos en la harina, para de esta forma poder hacer correcciones en los gráficos de control del proceso y así evitar elevadas pérdidas económicas por productos fuera de especificaciones, consumo de energía innecesario y emisiones atmosféricas (Información aportada por los especialistas de la empresa).

ANÁLISIS DE ESTUDIOS REALIZADOS

Para la cuantificación de los porcentajes de las fases minerales del clínker se han desarrollado métodos tales como [1] la Difracción de Rayos–X [6], químico–Cálculo Potencial de Bogue [7, 8] y microscopía óptica [3, 9,10,11,12,13,14,15]. Este último, se realiza mediante conteo manual de puntos [16] o análisis digital de Imágenes [17].

En Cuba (según los especialistas de la empresa CCSA) y el mundo el método clásico de cuantificar los porcentajes de los minerales del clínker es usando las ecuaciones propuestas por Bogué hace cerca de un siglo [1], conocidas como Cálculo Potencial de Bogue [7]. El cálculo se realiza a partir del conocimiento de los porcentajes de los principales óxidos que están presentes en el clínker. En su formulación estas ecuaciones asumen materias primas con pureza y reacción entre ellas del 100%, lo cual no es cierto para la mayoría de las cementeras, donde se tienen diferentes combinaciones de materias primas y procesos de clinkerización no totalmente controlados. Además el error se incrementa por la formación de compuestos menores y por la presencia de soluciones sólidas entre los componentes principales y menores [18].

[19], [4] y [18] muestran como algunos investigadores han concluido que los cálculos con las ecuaciones de Bogué generalmente subestiman el contenido de alita y sobreestiman el de belita y celita hasta en un 10% agregando, que el conteo de puntos mediante microscopía óptica puede producir un resultado más preciso para estas fases.

Autores como [20] identifican como una dificultad en el método óptico la cuantificación de los aluminatos (celitas) y ferroaluminatos (ferritas) presentes en la fase intersticial debido principalmente al tamaño tan pequeño de los cristales, los cuales pueden llegar incluso a ser amorfos. Ellos proponen para subsanar esta dificultad el empleo de Difracción de Rayos X Cuantitativa (QXRD) la cual puede ser difícil porque este es un material multifases y varios picos se superponen, sin embargo, ellos plantean que con los desarrollos del método de Rietveld se pueden minimizar o eliminar estos errores.

Aceptando que esta dificultad existe, es importante resaltar que una ventaja comparativa que tiene la microscopía óptica sobre estas otras técnicas es que además de cuantificar las fases permite ver las texturas y las alteraciones en los cristales como son: el tamaño de los cristales, distribución de las fases dentro de la muestra, clúster, retrogradaciones, zonaciones, maclas, etc. [4, 10, 14], y ahora con el desarrollo de sistemas de análisis digital de imágenes (ADI) esto se puede hacer en muy corto tiempo [17].

No obstante a lo planteado por [19], [4] y [18], la práctica ha demostrado que en la fábrica de cemento de Cienfuegos Cementos Cienfuegos S.A los cálculos con las ecuaciones de Bogué para cuantificar los porcentajes de los minerales del clínker brindan resultados suficientemente buenos (Información aportada por los especialistas de la empresa).

En el mundo, se han realizado variaciones a las ecuaciones de Bogue ya que cada cementera tiene materias primas y procesos de producción con características propias. Es por esto que autores como [21], [22] y [8] han propuesto modificaciones a dichas ecuaciones a partir de las composiciones químicas y mineralógicas para la dosificación y corrección de la pasta, pero ninguna de éstas ha recibido una aceptación general [1] mostrando que el problema es propio de cada planta.

Todas estas modificaciones han estado en función de la transformación de los coeficientes de las ecuaciones a partir del conocimiento de los porcentajes de los principales óxidos en el clínker, pero ninguna de ellas considera estas modificaciones a partir del conocimiento de los porcentajes de los principales óxidos en la harina.

En Colombia ya se conoce la existencia de una modificación de estas ecuaciones en una industria cementera con gran aceptación. Pero al igual que las anteriores esta modificación se realizó a partir del conocimiento de los porcentajes de los principales óxidos en el clínker.

REFORMULACIÓN DE LAS ECUACIONES DE BOGUE

El método empleado se organiza metodológicamente en dos etapas básicas:

1. Recopilación de la información y preparación de los datos.
2. Establecimiento de los modelos matemáticos para el cálculo de los porcentajes de las fases minerales del clínker.

Cada una de las etapas consta de un conjunto de pasos. La siguiente figura muestra las etapas con sus correspondientes pasos.

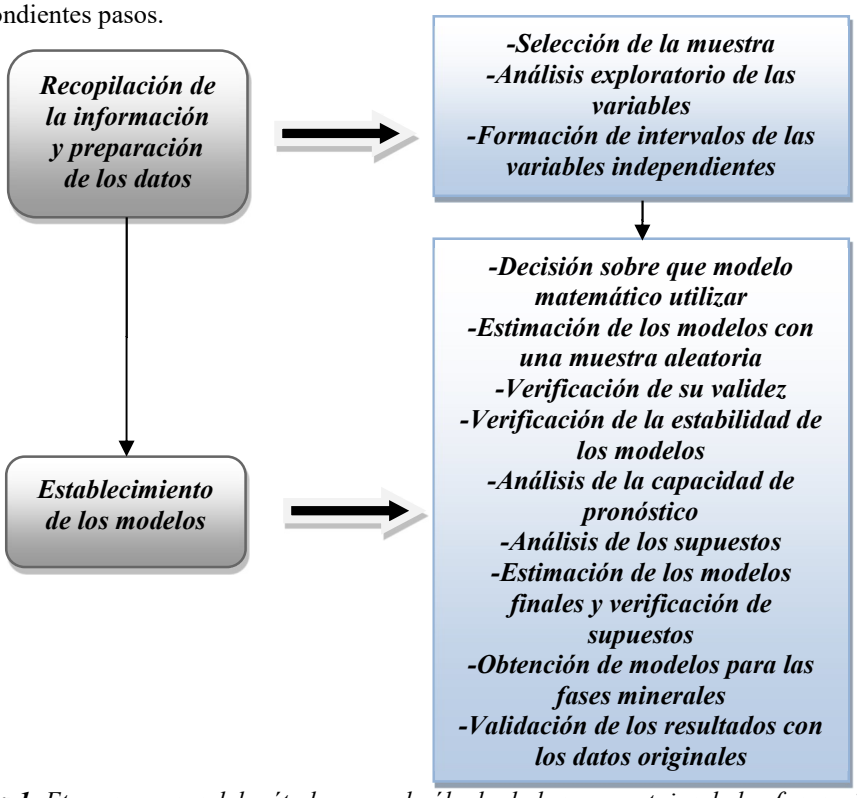


***Figura 1**. Etapas y pasos del método para el cálculo de los porcentajes de las fases minerales del clínker. Fuente: Elaboración propia.*

El hecho de que el método conste de las etapas mencionadas anteriormente se debe a lo siguiente:

Si se conoce que el porcentaje de cada óxido en el clínker depende del porcentaje del mismo óxido en la harina (Información aportada por los especialistas de la industria), determinando estas dependencias podría predecirse

el porcentaje de cada una de las fases minerales del clínker a partir del conocimiento de los porcentajes de los principales óxidos en la harina, por tanto se deben determinar las relaciones funcionales existentes entre los porcentajes de los óxidos en el clínker y en la harina.

Sea

$$CaO = f\left(CaO_H\right)(5)$$

la relación funcional entre el óxido de calcio en el clínker y el óxido de calcio en la harina, donde:

CaO es el óxido de calcio en el clínker y CaO_H es el óxido de calcio en la harina,

$$SiO_2 = f\left(SiO_{2H}\right)(6)$$

la relación funcional entre el dióxido de sílice en el clínker y el dióxido de sílice en la harina, donde:

SiO_2 es el dióxido de sílice en el clínker y SiO_{2H} es el dióxido de sílice en la harina,

$$Al_2O_3 = f\left(Al_2O_{3H}\right)(7)$$

la relación funcional entre el óxido de aluminio en el clínker y el óxido de aluminio en la harina, donde:

Al_2O_3 es el óxido de aluminio en el clínker y Al_2O_{3H} es el óxido de aluminio en la harina,

$$Fe_2O_3 = f\left(Fe_2O_{3H}\right)(8)$$

la relación funcional entre el óxido de hierro en el clínker y el óxido de hierro en la harina, donde:

Fe_2O_3 es el óxido de hierro en el clínker y Fe_2O_{3H} es el óxido de hierro en la harina y

$$SO_3 = f\left(SO_{3H}\right)(9)$$

la relación funcional entre el trióxido de azufre en el clínker y el trióxido de azufre en la harina, donde:

SO_3 es el trióxido de azufre en el clínker y SO_{3H} es el trióxido de azufre en la harina, si sustituimos (5), (6), (7), (8) y (9) en(1), (2), (3), (4) y (5), obtenemos:

$$C3S = 4.071f\left(C_aOH\right) - \begin{pmatrix} 7.6f\left(S_iO_2H\right) + 6.718f\left(Al_2O_3H\right) + \\ +1.43f\left(Fe_2O_3H\right) + 2.852f\left(SO_3H\right) \end{pmatrix}(10)$$

$$C2S = 2.867f\left(S_iO_2H\right) - 0.7544C_3S\,(11)$$

$$C3A = 2.65f\left(Al_2O_3H\right) - 1.692f\left(Fe_2O_3H\right)(12)$$

$$C4AF = 3.043f\left(Fe_2O_3H\right)(13)$$

y con estas ecuaciones (10), (11), (12) y (13) puede predecirse el porcentaje de cada una de las fases minerales del clínker a partir del conocimiento de los porcentajes de los principales óxidos en la harina, pero para la obtención de las ecuaciones anteriores se debe recopilar información y preparar los datos relacionados con los porcentajes de los óxidos tanto en la harina como en el clínker y establecer los modelos (10), (11), (12) y (13) después de hallar las relaciones funcionales (5), (6), (7), (8) y (9).

APLICACIÓN DEL MÉTODO

Los resultados de la aplicación del método se presentan a continuación sobre la base de los resultados obtenidos a partir de la ejecución de cada uno de los pasos de cada etapa.

- **Resultados de la primera etapa**

En la investigación se trabaja con toda la información generada en el período de enero de 2013 a abril de 2013 el cual se considera representativo del proceso productivo y de la calidad de los materiales que se utilizan. Los

datos se corresponden con los resultados de los análisis químicos que se realizan en la fábrica y están contenidos en una base de datos en formato .xls que contiene un total de 1117 observaciones de cada una de las variables mencionadas con anterioridad.

Como resultado del análisis exploratorio de las variables los estadísticos descriptivos (media, desviación típica, mínimo y máximo) y los intervalos de confianza para la media de las variables que se utilizan en el estudio se contrastan con los requerimientos técnicos de estas variables en la producción del clínker observándose que los intervalos de confianza para la media de todas las variables son un subconjunto del intervalo que garantiza los requisitos de calidad del cliente y además se aprecia que el valor típico de cada una de las variables de acuerdo con los requisitos del cliente no está incluido en el intervalo de confianza de su variable correspondiente.

A partir de los gráficos de tallos y hojas y los diagramas de cajas de las variables que se utilizan en el estudio, se excluyen de los análisis para obtener los modelos dados por las ecuaciones (5), (6), (7), (8) y (9) los valores atípicos encontrados para cada una de las variables, eliminándose 24 observaciones de la variable SiO_{2H}, 43 de la variable Al_2O_{3H}, 10 de la variable CaO_H, 8 de la variable SO_{3H}, 41 de la variable SiO_2, 30 de la variable Al_2O_3, 9 de la variable Fe_2O_3, 38 de la variable CaO y 37 de la variable SO_3. La observación de los gráficos de tallos y hojas permite además sospechar una posible normalidad de las variables en estudio.

En el estudio de las correlaciones bivariadas de Pearson y los diagramas de dispersión entre las variables correspondientes de las variables que se utilizan en el estudio se puede apreciar que los coeficientes de correlación de Pearson entre las variables dependientes e independientes son estadísticamente significativos pero muy débiles, mientras que los gráficos de dispersión entre estas variables no muestran otras relaciones funcionales además de estas relaciones lineales, comportamiento que no favorece la obtención de modelos que permitan estimar con exactitud los porcentajes de los óxidos en el clínker a partir de los porcentajes de los óxidos en la harina, por tanto y en base al comportamiento de los óxidos en el clínker que muestran que cuando en la harina se producen incrementos pequeños, menores o iguales a 0.1% para la variable CaO_H, 0.02% para las variables SO_{3H}, Al_2O_{3H} y Fe_2O_{3H} y en el caso del SiO_{2H} inferiores o iguales a 0.05%, estos incrementos no inciden significativamente en los porcentajes de estos óxidos en el clínker, se procede a formar intervalos de amplitud menor o igual a 0.1 para la variable CaO_H, menor o igual a 0.02 para las variables SO_{3H}, Al_2O_{3H} y Fe_2O_{3H} y menor o igual a 0.05 para la variable SiO_{2H} y a tomar como valores para las variables independientes los puntos medios de cada uno de estos intervalos y como valores correspondientes de las variables dependientes los puntos medios de los intervalos correspondientes y así se formaron 22 intervalos para la variable CaO_H, 16 intervalos para las variables SiO_{2H} y SO_{3H}, 25 intervalos para la variable Al_2O_{3H} y 23 intervalos para la variable Fe_2O_{3H} y se crearon las variables:

$MediaCaOCAT$: que representa la media del porcentaje de óxido de calcio en el clínker para cada intervalo de la variable CaO_H,

$MediaCaOHCAT$: que representa la media del porcentaje de óxido de calcio en la harina para cada intervalo de la variable CaO_H,

$MediaSiO_2CAT$: que representa la media del porcentaje de dióxido de sílice en el clínker para cada intervalo de la variable SiO_{2H},

$MediaSiO_2HCAT$ que representa la media del porcentaje de dióxido de sílice en la harina SiO2H para cada intervalo de la variable SiO_{2H},

$MediaAl_2O_3CAT$ que representa la media del porcentaje de óxido de aluminio en el clínker para cada intervalo de la variable Al_2O_{3H},

$MediaAl_2O_3HCAT$ que representa la media del porcentaje de óxido de aluminio en la harina para cada intervalo de la variable Al_2O_{3H},

$MediaFe_2O_3CAT$ que representa la media del porcentaje de óxido de hierro en el clínker para cada intervalo de la variable Fe_2O_{3H},

$MediaFe_2O_3HCAT$ que representa la media del porcentaje de óxido de hierro en la harina para cada intervalo de la variable Fe_2O_{3H},

$MediaSO_3CAT$ que representa la media del trióxido de azufre en el clínker para cada intervalo de la variable SO_{3H} y

$MediaSO_3HCAT$ que representa la media del porcentaje de trióxido de azufre en la harina para cada intervalo de la variable SO_{3H}.

- **Resultados de la segunda etapa**

Dado que en el problema tratado no se puede obtener un ajuste exacto a todos los puntos ya que existen valores de las variables independientes para los cuales se tienen más de un valor de las variables dependientes se decide realizar un análisis de regresión, más específicamente un análisis de regresión lineal simple para estimar los modelos:

$$CaO_i = \beta_{11} + \beta_{21} CaOH_i + \varepsilon_i \quad (14)$$

$$SiO_{2_i} = \beta_{11} + \beta_{21} SiO_2H_i + \varepsilon_i \quad (15)$$

$$SO_{3_i} = \beta_{11} + \beta_{21} SO_3H_i + \varepsilon_i \quad (16)$$

$$Al_2O_{3_i} = \beta_{11} + \beta_{21} Al_2O_3H_i + \varepsilon_i \quad (17))$$

$$Fe_2O_{3_i} = \beta_{11} + \beta_{21} Fe_2O_3H_i + \varepsilon_i \quad (18)$$

La estimación de los modelos (14), (15), (16), (17) y (18) con una muestra aleatoria del 50% de los datos, y con las variables *MediaCaOCAT*, *MediaCaOHCAT*, $MediaSiO_2CAT$, $MediaSiO_2HCAT$, $MediaAl_2O_3CAT$, $MediaAl_2O_3HCAT$, $MediaFe_2O_3CAT$, $MediaFe_2O_3HCAT$, $MediaSO_3CAT$, $MediaSO_3HCAT$ obtenidas por los intervalos formados para cada variable, arrojó como resultado la obtención de los modelos dados por las ecuaciones siguientes:

$$CaO = 60.812 + 0.117 CaOH \quad (19)$$

$$SiO_2 = 18.478 + 0.194 SiO_2H \quad (20)$$

$$SO_3 = 1.460 + 0.510 SO_3H \quad (21)$$

$$Al_2O_3 = 2.236 + 0.884 Al_2O_3H \quad (22)$$

$$Fe_2O_3 = 1.221 + 1.044 Fe_2O_3H \quad (23)$$

A los modelos (19), (20), (21), (22) y (23) se le comprueba la validez con la prueba *T* de Student, la cual arroja resultados satisfactorios, ya que se puede constatar que la significación del estadígrafo es menor en todos los casos que el nivel de significación $\alpha = 0.05$ y por tanto se concluye que todos los modelos son válidos.

Con el 50% de los datos no seleccionados para la obtención de los modelos (19), (20), (21), (22) y (23) se obtienen los modelos:

$$CaO = 61.171 + 0.108 CaOH \quad (24)$$

$$SiO_2 = 19.247 + 0.138 SiO_2H \quad (25)$$

$$SO_3 = 1.430 + 0.569 SO_3H \quad (26)$$

$$Al_2O_3 = 1.920 + 0.988 Al_2O_3H \quad (27)$$

$$Fe_2O_3 = 1.593 + 0.874 Fe_2O_3H \quad (28)$$

y se compara el parámetro correspondiente a la variable independiente en dichos modelos contra el valor del parámetro obtenido en los primeros modelos estimados (19), (20), (21), (22) y (23) utilizando la prueba T de Student la cual arroja como resultado que los modelos no cambian significativamente con el cambio de la muestra, es decir son estables en la población en estudio.

Para verificar la capacidad de pronóstico de los modelos (19), (20), (21), (22) y (23) se aplica la prueba *T* de Student en 2 poblaciones mediante un diseño de muestras pareadas, donde una población está constituida por los valores de la variable dependiente en los datos no utilizados en la estimación de los modelos y la segunda población está dada por los pronósticos de estos datos realizados a partir de los modelos estimados, la hipótesis nula en esta prueba es que las medias en ambas poblaciones son iguales lo que se reduce a verificar la hipótesis nula de que la diferencia de las medias en ambas poblaciones es cero o lo que es lo mismo, que la media de los errores es cero. Esta prueba se aplica para cada uno de los modelos y arroja como resultado que los modelos tienen una buena capacidad de pronóstico, por tanto se puede concluir que todos los modelos estimados pronostican adecuadamente, además en todos los casos se puede apreciar que los intervalos de confianza para la media de los pronósticos de cada uno de los modelos son un subconjunto de los intervalos que garantizan los requisitos del cliente.

El análisis de los supuestos de cada uno de los modelos (19), (20), (21), (22) y (23) también arroja resultados satisfactorios.

Una vez realizado el análisis de los supuestos de los modelos dados por las ecuaciones (19), (20), (21), (22) y (23) se obtienen los modelos:

$$CaO = 61.642 + 0.097 CaOH \quad (29)$$

$$SiO_2 = 18.791 + 0.171 SiO_2H \quad (30)$$

$$SO_3 = 1.444 + 0.541 SO_3H \quad (31)$$

$$Al_2O_3 = 2.162 + 0.908 Al_2O_3H \quad (32)$$

$$Fe_2O_3 = 1.303 + 1.007 Fe_2O_3H \quad (33)$$

con el 100% de los datos. Estos modelos (29), (30), (31), (32) y (33) representan las relaciones funcionales (5), (6), (7), (8) y (9). A estos modelos también se les realiza un análisis de los supuestos, obteniéndose en todos los casos resultados satisfactorios.

A partir de la sustitución de los modelos (29), (30), (31), (32) y (33) en las ecuaciones (1), (2), (3) y (4) se obtienen los modelos:

$$C3S = 87.627088 + 0.394887 C_aOH - \begin{pmatrix} 1.2996 SiO_2H + 6.099944 Al_2O_3H + \\ + 1.44001 Fe_2O_3H + 1.542932 SO_3H \end{pmatrix} \quad (34)$$

$$C2S = 53.873797 + 0.490257 S_iO_2H - 0.7544 C_3S \quad (35)$$

$$C3A = 3.524624 + 2.4062 Al_2O_3H - 1.703844 Fe_2O_3H \quad (36)$$

$$C4AF = 3.965029 + 3.064301 Fe_2O_3H \quad (37)$$

los cuales representan las ecuaciones (10), (11), (12) y (13) y permiten calcular los porcentajes de las fases minerales del clínker a partir del conocimiento de los porcentajes de los óxidos en la harina.

La validación de los modelos (34), (35), (36) y (37) se realiza mediante la verificación de la capacidad de pronóstico con una muestra adicional de 400 observaciones no empleadas en la estimación aplicando la prueba *T* de Student a los errores dados por la diferencia de los pronósticos de estos modelos con los valores reales de la variable dependiente. Esta prueba arrojó resultados satisfactorios.

El método estudiado permite reducir daños al Medio Ambiente puesto que en el proceso productivo se eliminan producciones por productos fuera de especificaciones dadas por los clientes.

En resumen, al aplicar el método matemático para el cálculo de las fases minerales del clínker se logra la obtención de 5 modelos que permiten predecir los porcentajes de los principales óxidos en el clínker a partir del conocimiento de los porcentajes de estos óxidos en la harina, estos modelos son:

$$CaO = 61.642 + 0.097 CaOH \quad (29)$$

$$SiO_2 = 18.791 + 0.171 SiO_2H \quad (30)$$

$$SO_3 = 1.444 + 0.541 SO_3H \quad (31)$$

$$Al_2O_3 = 2.162 + 0.908 Al_2O_3H \quad (32)$$

$$Fe_2O_3 = 1.303 + 1.007 Fe_2O_3H \quad (33)$$

y mediante la sustitución de estos modelos en las Ecuaciones de Bogue se obtiene un conjunto de ecuaciones que permiten predecir los porcentajes de las fases minerales del clínker también a partir del conocimiento de los porcentajes de los óxidos en la harina. Estas ecuaciones son:

$$C3S = 87.627088 + 0.394887 C_aOH - \begin{pmatrix} 1.2996 SiO_2H + 6.099944 Al_2O_3H + \\ + 1.44001 Fe_2O_3H + 1.542932 SO_3H \end{pmatrix} \quad (34)$$

$$C2S = 53.873797 + 0.490257 S_iO_2H - 0.7544 C_3S \quad (35)$$

$$C3A = 3.524624 + 2.4062 Al_2O_3H - 1.703844 Fe_2O_3H \quad (36)$$

$$C4AF = 3.965029 + 3.064301 Fe_2O_3H \quad (37)$$

CONCLUSIONES

Los métodos que se usan en la actualidad para la cuantificación de los porcentajes de las fases minerales del clínker no responden a las necesidades actuales de los especialistas de la Empresa Mixta Cementos Cienfuegos S.A. Lo más utilizado para la cuantificación de los mencionados porcentajes son las Ecuaciones de Bogue, las cuales usan para el cálculo de los porcentajes de las fases minerales los porcentajes de los principales óxidos en el clínker y no en la harina, por tanto en esta investigación se establece un método matemático para el cálculo de los porcentajes de las fases minerales del clínker sustentado en 5 modelos que permiten realizar dichos cálculos a partir del conocimiento de los porcentajes de los principales óxidos en la harina. El método se valida comparando los resultados de la aplicación del mismo con los resultados obtenidos por la empresa por análisis químico e influye en la reducción de los daños al Medio Ambiente derivados del proceso de fabricación del cemento debido a que se elimina la necesidad de realizar nuevas producciones por productos fuera de especificaciones.

RECOMENDACIONES

Que se continúe utilizando el método propuesto en la Empresa Mixta Cementos Cienfuegos S.A
Aplicar el método en otras industrias cementeras del país para así evitar pérdidas económicas por producciones de baja calidad y daños innecesarios al medio ambiente.
Que se divulgue la experiencia de esta investigación otras empresas cementeras de otros países.

REFERENCIAS

[1] J. I. Tobón, "Replanteamiento de las Ecuaciones de Bogue en el cálculo mineralógico del clínker para una cementera colombiana," 2006.

[2] A. Agostini, "Características Medioambientales y Riesgos en la salud por los Materiales de Construcción en las Edificaciones," 2007.

[3] Holderbank, "Química y mineralogía de las materias primas del cemento: Influencia de las propiedades de las materias primas en el proceso de fabricación del cemento," in *Curso del Cemento* Bogotá, 1975, p. 46.

[4] F. Glasser, "The burning of Portland cement," 1998.

[5] W. H. Duda, *Manual tecnológico del cemento*, 1997.

[6] A. C1365-98, "Standard test method for determination of the proportion of phases in Portland clinker using X-ray diffraction analysis." vol. 4, 1998.

[7] A. C150-94, "Standard Specification for Portland Cement. ANEXO A1. Calculation of potential cement phase composition." vol. 4, 1994.

[8] M. Clark, "Bogue vs Chromy," *International cement review,* 2002.

[9] F. Calderón, "Resumen de microscopía de clinker Medellín," p. 16, 1977.

[10] E. Fundal, "Microscopy of cement raw mix and clinker," F.L.Smidth, Copenhagen 25, 1979.

[11] J. R. Camara, "Analise microscopica de clinker," Sao Paulo 16, 1988.

[12] L. F. Rodríguez, "Análisis microscópico del clinker de cemento portland," Medellín 53, 1991.

[13] N. Arenas, "Análisis químico instrumental en la industria del cemento. Observación microscópica cualitativa del clinker del cemento portland " 1997.

[14] D. H. Campbell, "Microscopical examination and interpretation of portland cement and clinker," *PORTLAND CEMENT ASSOCIATION,* vol. 2, 1999.

[15] J. I. Tobón, "Caracterización petrográfica de algunos clinkeres Colombianos," 2001.

[16] A. C1356M-96, "Standard test method for quantitative of phases in Portland cement clinker by microscopical point-count procedure." vol. 4, 1996.

[17] J. M. García-Márquez, "Automatic quantification of phases and mechanical characterization of materials base don Pórtland clinker modified with silica and alumina additions," *Journal of Materials Proccessing Technology,* p. 4, 2003.

[18] D. Lawrence, "The constitution and specification of portland cements," in *Lea´s chemistry of cement and concrete*, 4 ed, 1998, p. 64.

[19] H. F. W. Taylor, *Cement Chemistry*, 2 ed., 1997.

[20] A. Crumbie, "Where is the iron? Clinker microanálisis with XRD Rietveld, optical microscopy/point counting, Bogue and SEMEDS techniques.Cement and Concrete and Research," p. 6, 2006.

[21] E. Marciano, "Estudio comparativo entre dos resultados obtenidos microscopicamente e pelo metodo de bogue e suas implicacoes," Sao Paulo1983.

[22] H. F. W. Taylor, "Modification of the Bogue calculation," *Advances in Cement Research,* 1989.

Capítulo 7 pp 97-119

CARACTERIZACIÓN DE LOS INDICADORES AGRARIOS DE LA PRODUCCIÓN DE LIMÓN PERSA EN EL MUNICIPIO DE MARTÍNEZ DE LA TORRE, VERACRUZ, MÉXICO

Ignacio Caamal Cauich[1], Verna Gricel Pat Fernández[1] y José Félix García Rodríguez[2]

[1]Universidad Autónoma Chapingo, México.
[2]Universidad Juárez Autónoma de Tabasco, México. E-mail: jfgr55@hotmail.com

RESUMEN

Los limones aportan la mayor proporción del valor de la producción de la rama de los cítricos en México, seguidos por las naranjas. Las principales variedades de limones, por la superficie cosechada y volumen de la producción generada en México, son el limón persa y el limón mexicano. Los principales estados productores de limón persa en México son Veracruz, Oaxaca, Jalisco y Tabasco. Se calcularon los indicadores agrarios y agrícolas de la producción del limón persa en el municipio de Martínez de la Torre, Veracruz, con base en los coeficientes de participación de las variables de producción. El principal estado productor de limón persa en México es Veracruz (54%), en donde la mayor parte de la producción se ubica en el Distrito de Desarrollo Rural de Martínez de la Torre (79%), especialmente en el municipio de Martínez de la Torre, con un 44% de la producción en el Distrito. La superficie cosechada promedio por productor es de 5.3 ha; cerca del 80% de los productores tienen una superficie de 3 a 6 ha y una superficie total de alrededor del 45%; mientras que alrededor de 12% de los productores tienen de 6.1 a 10 ha y poseen alrededor del 20% de la superficie cosechada; y, finalmente, cerca del 10% de los productores tienen más de 9 ha y cuentan con cerca de 35% de la superficie total. Los datos reflejan una gran diferenciación entre los productores, donde la inmensa mayoría tiene superficies pequeñas y una minoría posee grandes superficies de limón persa.

Palabras clave: superficie, rendimiento, producción.

CHARACTERIZATION OF AGRICULTURAL INDICATORS OF PERSIAN LIME PRODUCTION IN THE MUNICIPALITY OF MARTÍNEZ DE LA TORRE, VERACRUZ, MEXICO

ABSTRACT

Lemons contribute the largest proportion of the value of citrus production in Mexico, followed by oranges. The main varieties of lemons, due to the area harvested and volume of production generated in Mexico, are the Persian lime and the Mexican lemon. The main Persian lime producing states in Mexico are Veracruz, Oaxaca, Jalisco and Tabasco. The agrarian and agricultural indicators of the production of the Persian lime in the municipality of Martínez de la Torre, Veracruz, were calculated based on the participation coefficients of the production variables. The main Persian lemon producing state in Mexico is Veracruz (54%), where most of the production is located in the Rural Development District of Martínez de la Torre (79%), especially in the municipality of Martínez de la Torre, with 44% of the production in the District. The average harvested area per producer is 5.3 ha; about 80% of producers have an area of 3 to 6 ha and a total area of about 45%; while about 12% of producers have 6.1 to 10 ha and own about 20% of the harvested area; and, finally, about 10% of

[1] E-mail: icaamal82@yahoo.com.mx

the producers have more than 9 ha and have about 35% of the total area. The data reflect a great differentiation between producers, where the vast majority have small areas and a minority have large areas of Persian lemon.

Key words: surface, yield, production.

1. INTRODUCCIÓN

Antecedentes

El limón es uno de los cítricos más importantes en el mundo, por el valor de la producción que genera, tiene sus orígenes en el continente asiático, de acuerdo con los aportes de la literatura relacionada con el cultivo (Morton, 1987; ASERCA, 1996). Fue introducido al norte de África y a Europa, principalmente a España, por los árabes, finalmente de España el limón llega a México, con la colonización de México por los españoles (ASERCA, 1996).

Las principales variedades de limas y limones son el *Citrus limón*, *Citrus latifolia*, *Citrus aurantifolia* y *Citrus limetta*, siendo la variedad *Citrus limón* Eureka la más importante en el mundo, seguido por la variedad *Citrus latifolia*, que también tiene importancia comercial, sin embargo, esta variedad no es propiamente un limón como tal, sino una variedad de limas ácidas, razón por la cual no tiene semillas (CCI, 2000).

El cultivo del limón persa se realiza en una franja que inicia desde el Ecuador hasta los 40 grados de latitud Norte y Sur, dentro de la cual predominan los climas tropicales y subtropicales. Así mismo, el limón persa se puede desarrollar en lugares con temporadas de lluvias en verano, teniendo un promedio de alrededor de 880 mm anuales, con temperaturas que varían de 1 a 40 grados centígrados y prospera en terrenos de textura limo-arcillosa (ASERCA, 1995).

El limón persa (*Citrus latifolia* L.), también conocido como “limón sin semilla”, “lima de Persia” o “lima de Tahití”, es un fruto de forma oblongo a ovoide, con una papila terminal ancha no muy pronunciada, de 3.8 a 6.6 cm de largo e incluso mayor, de color amarillo brillante al madurar, con ligeras rugosidades, con ocho o diez segmentos, ácido, de producción media y preferido por su buena calidad para exportar como fruta fresca (ASERCA, 1995). El limón persa es un cultivo triploide, por lo que el fruto carece de semillas, además de que el árbol tiene una menor cantidad de espinas en comparación con el limón mexicano, lo que facilita el proceso de cosecha (Khan *et al.*, 2017).

El fruto del limón persa tiene diversos usos, destacando el consumo en fresco para la utilización del jugo, principalmente como condimento en alimentos y bebidas; la producción de materia prima, extrayendo de la cáscara el aceite esencial, que es ampliamente utilizado en la formulación de diferentes productos en la industria alimentaria, farmacéutica, cosmética y de productos de limpieza, tales como aromatizantes y desinfectantes (Montero, 2009); además, el fruto es aprovechado en la industria para la preparación de concentrados, ácido cítrico y pectinas, mientras que la pulpa o bagazo sirve para la alimentación de ganado (IICA *et al.*, 2004).

De acuerdo con datos del Departamento de Agricultura de Estados Unidos (USDA), el limón persa, en una porción de 67 g (una fruta), aporta 20.1 kcal, 1.3 mg de sodio y 1.1 g de azucares, y como porcentaje de un valor diario con base en una dieta de 2,000 calorías, proporciona 7.0 g de carbohidratos totales (2.6%), 1.9 g de fibra dietética (6.7%), 19.5 mg de Vitamina C (21.7%), 22.1 mg de Calcio (1.7%), 0.4 mg de Hierro (2.2%) y 33.5 µg de Vitamina A (3.7%), entre otros (USDA, 2019).

Importancia económica

La superficie cosechada de cítricos fue de alrededor de 9.9 millones de hectáreas (ha), las cuales generaron una producción de aproximadamente 158.0 millones de toneladas (ton) a nivel mundial. El principal cultivo citrícola en el mundo es la naranja con 4,060,129 ha (41.0%) de la superficie cosechada y 78,699,604 ton (49.8%) de la producción total de cítricos, le siguen el grupo de las tangerinas, mandarinas, entre otros, con 2,756,887 ha (27.9%) de la superficie cosechada y 35,444,080 ton (22.4%) de la producción; y en tercer lugar los limones y limas con 1,226,617 ha (12.4%) de la superficie cosechada y 20,049,630 ton (12.7%) de la producción; el resto se integra con toronjas y pomelos con 346,191 ha (3.5%) de la superficie cosechada y 9,289,462 ton (5.9%) de la producción y otros cítricos no clasificados precedentemente con 1,508,639 ha (15.2%) de la superficie cosechada y 14,496,484 ton (9.2%) de la producción (FAOSTAT, 2021 y Figura 1).

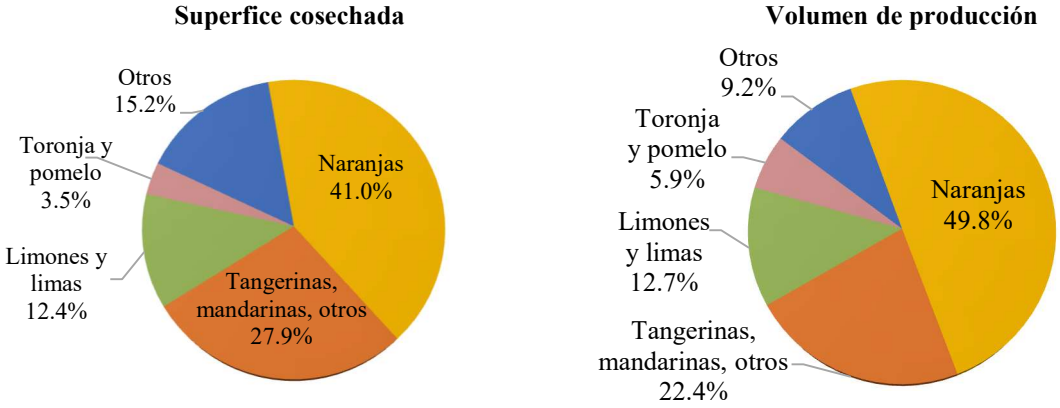


Figura 1. Distribución de la superficie cosechada y volumen de producción cítricos a nivel mundial, 2019. Fuente: Elaborado con datos de FAOSTAT (2021).

En el año de 2019 se registraron en el mundo 9,898,463 hectáreas cosechadas de cítricos, con una producción de 157,979,260 toneladas, de las cuales 1,226,617 ha (12.4%) corresponden a limones y limas, con una producción de 20,049,630 toneladas (12.7%) y el resto a otros cítricos (FAOSTAT, 2021).

Los principales países productores de limones y limas en el mundo son la India con 305,000 ha (24.9%) de la superficie cosechada y 3,482,000 ton (17.4%) de la producción; México con 183,575 ha (15.0%) de la superficie cosechada y 2,701,828 ton (13.5%) de la producción; China C. con 128,823 ha (10.5%) de la superficie cosechada y 2,666,082 ton (13.3%) de la producción; Argentina con 57,541 ha (4.7%) de la superficie cosechada y 1,904,765 ton (9.5%) de la producción; Brasil con 56,491 ha (4.6%) de la superficie cosechada y 1,511,185 ton (7.5%) de la producción; entre otros, los países restantes con 495,187 ha (40.4%) de la superficie cosechada y 7,783,770 ton (38.8%) de la producción en México (FAOSTAT, 2021 y Figura 2).

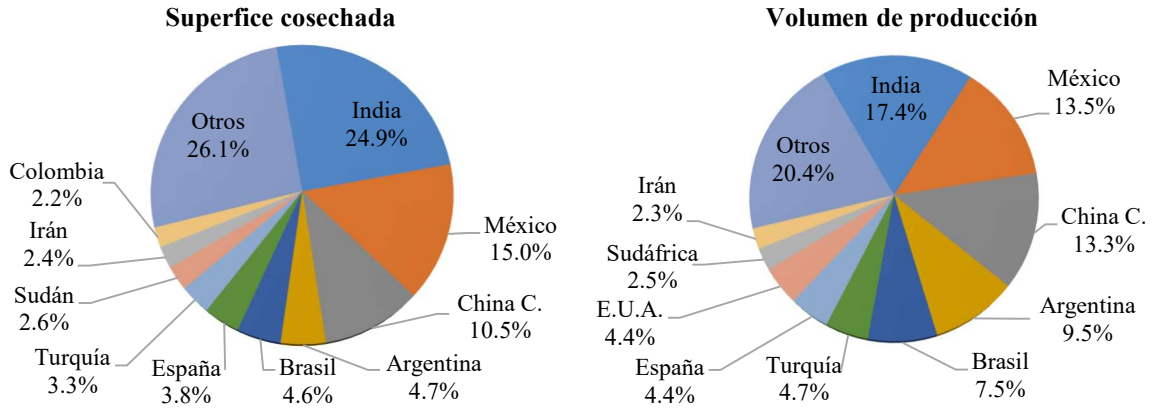


Figura 2. Distribución de la superficie cosechada y volumen de producción de limones y limas a nivel mundial, 2019.
Fuente: Elaborado con datos de FAOSTAT (2021).

La citricultura es una de las actividades económicas más importantes dentro del sector agrícola en México, ya que ocupa 617,254 ha, 36.1% de la superficie destinada a frutales y aporta 8,686,667 ton, alrededor del 36.4% del volumen de producción de frutales (SIACON, 2021 y Figura 3).

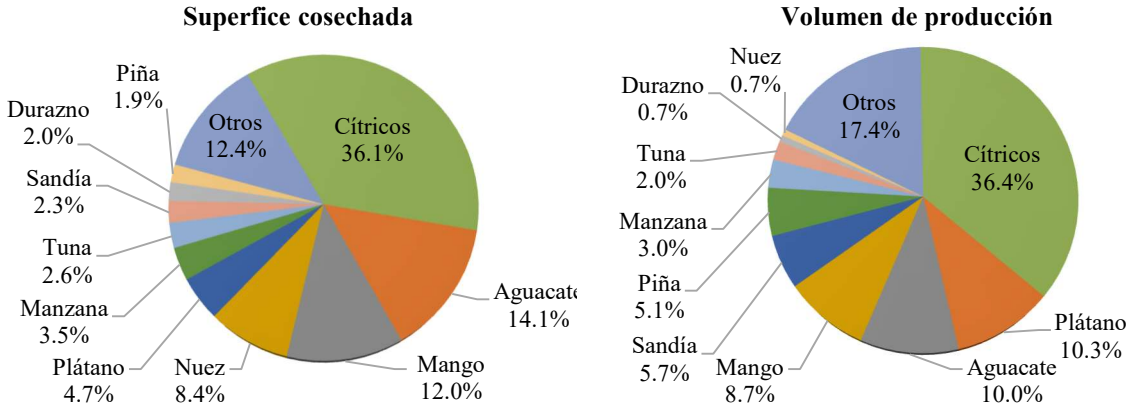


Figura 3. Distribución de la superficie cosechada de frutales en México, 2020.
Fuente: Elaborado con datos de SIACON (2021).

Los principales limones y limas que se producen en México son el limón persa con 94,595 ha (50.5%) de la superficie cosechada y 1,447,416 ton (50.3%) de la producción; limón mexicano con 81,728 ha (43.6%) de la superficie cosechada y 1,221,459 ton (42.4%) de la producción; limón italiano con 7,142 ha (3.8%) de la superficie cosechada y 140,748 ton (4.9%) de la producción; lima con 2,364 ha (1.3%) de la superficie cosechada y 27,596 ton (1.0%) de la producción; y otros con 1,651 ha (0.9%) de la superficie cosechada y 41,804 ton (1.5%) de la producción (SIACON, 2021 y Figura 4).

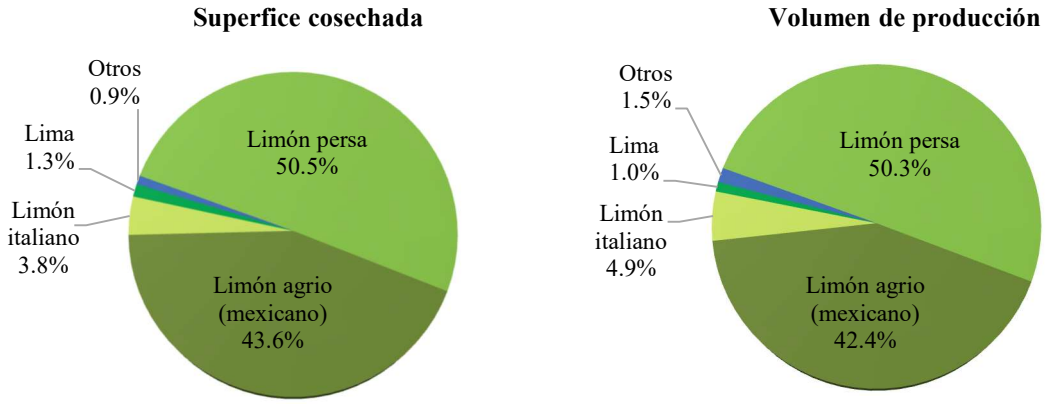


Figura 4. Distribución de la superficie cosechada y volumen de producción de limones y limas en México, 2020.
Fuente: Elaborado con datos de SIACON (2021).

Los principales estados productores de limón persa en México son Veracruz con 46,959 ha (49.6%) de la superficie cosechada y 785,668 ton (54.3%) de la producción generada; Oaxaca con 14,350 ha (15.2%) de la superficie cosechada y 198,586 ton (13.7%) de la producción aportada; Jalisco con 4,996 ha (5.3%) de la superficie cosechada y 92,973 ton (6.4%) de la producción creada; Tabasco con 7,223 ha (7.6%) de la superficie cosechada y 86,097 ton (5.9%) de la producción generada; Yucatán con 3,683 ha (3.9%) de la superficie cosechada y 77,205 ton (5.3%) de la producción creada; entre otros. Los cinco principales estados señalados poseen más del 80% de la superficie cosechada y aportan más del 85% de la producción generada (SIAP, 2021 y Figura 5).

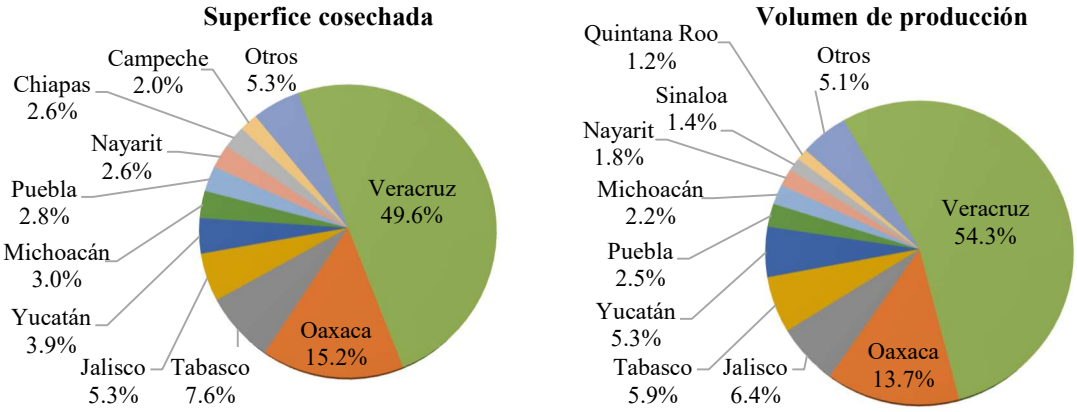


Figura 5. Distribución de la superficie cosechada y volumen de producción de limón persa en México, 2019.
Fuente: Elaborado con datos de SIACON (2021).

Como en la mayoría de los cultivos en México, la producción del limón persa se concentra mayormente en periodo de temporal con alrededor de 68,106 ha, un 68% de la superficie sembrada, principalmente en los estados de Veracruz, Oaxaca y Tabasco. La superficie cosechada de riego es de 32,343 ha, el 32% de la superficie sembrada total de 100,449 ha, ubicada principalmente en Veracruz, Jalisco y Yucatán (SIACON, 2021 y Figura 6).

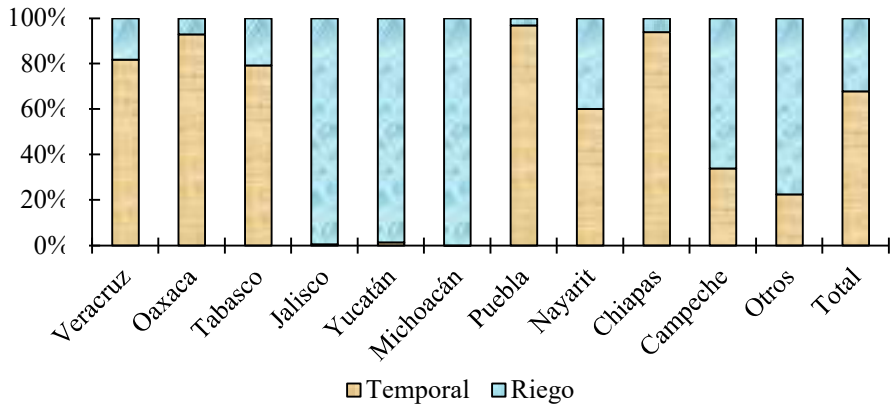


Figura 6. Superficie de riego y temporal de limón persa en México, 2020 (%).
Fuente: Elaborado con datos de SIACON (2021).

La principal zona productora de limón persa en el estado de Veracruz es el Distrito de Desarrollo Rural de Martínez de la Torre, que concentró 35,010 ha, las cuales corresponden al 37.0% de la superficie cosechada de limón persa en México, de 100,449 ha y alrededor del 75% de la superficie estatal cultivada de 47,809 ha, con un rendimiento promedio de 17.6 ton/ha, el más alto de los Distritos de Desarrollo Rural del estado. La producción generada en el Distrito de Desarrollo Rural de Martínez de la Torre fue de 617,687 toneladas, cerca del 43% de la producción nacional de 1,447,416 toneladas y alrededor del 79% de la producción estatal de 785,668 toneladas (SIACON, 2021 y Figura 7).

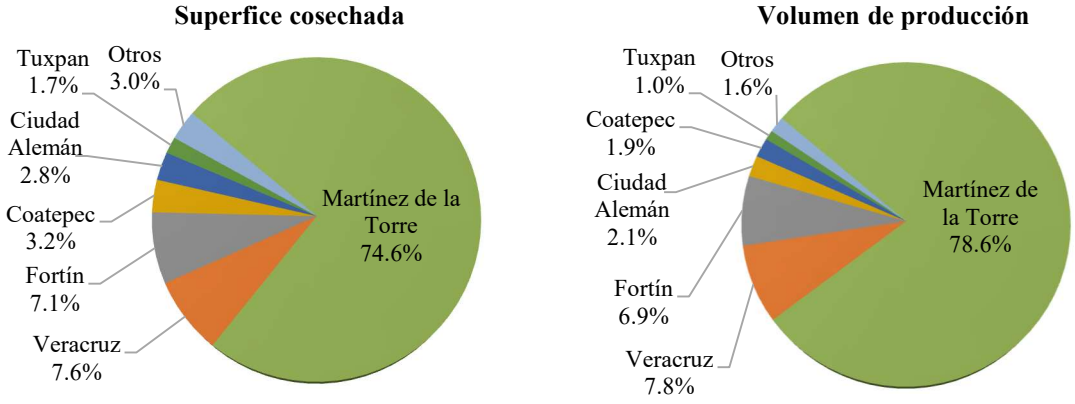


Figura 7. Distribución de la superficie cosechada y volumen de producción de limón persa en el estado de Veracruz, principales DDR, 2020.
Fuente: Elaborado con datos de SIACON (2021).

Debido al gran impulso que han tenido los cítricos, especialmente en la producción de limón persa, en el municipio de Martínez de la Torre se ha desarrollado una fuerte infraestructura para la comercialización, se cuenta con más de 40 empacadoras de limón persa, dos plazas importantes de subasta de naranja, limón, toronja y otras frutas que se cultivan en la región, que generan un alto número de empleos a la población, debido a que

se requiere de una gran cantidad de jornales para la producción, empacado y operación de las subasta de limón persa, además el cultivo es generador de divisas, debido a que está orientado principalmente a la exportación (CONCITVER, 2012).

Almaguer y Ayala (2014), realizaron un diagnóstico y caracterización de la producción de limón persa con base en las variables superficie cultivada, rendimiento, edad de las plantaciones, poda y fertilización. Así mismo, Flores *et al*. (2014) realizan la caracterización de la competitividad con base en las variables ingresos, calidad, fertilización, humedad, podas, organización y agente y lugar de venta. En este trabajo se retoman las variables señaladas y se calculan los indicadores correspondientes.

El objetivo del trabajo fue realizar la caracterización de los indicadores agrarios de la producción de limón persa en el municipio de Martínez de la Torre, Veracruz, con la finalidad de identificar los problemas sociales y agrarios de la producción de limón persa. La hipótesis del trabajo plantea que la mayoría de los productores de limón persa son pequeños por el tamaño de la superficie, son grandes de edad, tienen bajos niveles de educación, obtienen rendimientos cercanos a la media nacional y utilizan una buena tecnología.

2. MATERIALES Y MÉTODOS

Localización del área de estudio

El municipio de Martínez de la Torre se encuentra ubicado en la zona Norte del estado de Veracruz, en las coordenadas 20° 04' de latitud Norte y 97° 04' de longitud Oeste, a una altura de 77 metros sobre el nivel del mar. Limita al Norte con el municipio de Tecolutla, al Este con los municipios de San Rafael y Misantla, al Sur con los municipios de Atzalan y Tlapacoyan y al Oeste con el municipio de Papantla y el estado de Puebla (INAFED, 2018 y Figura 8).

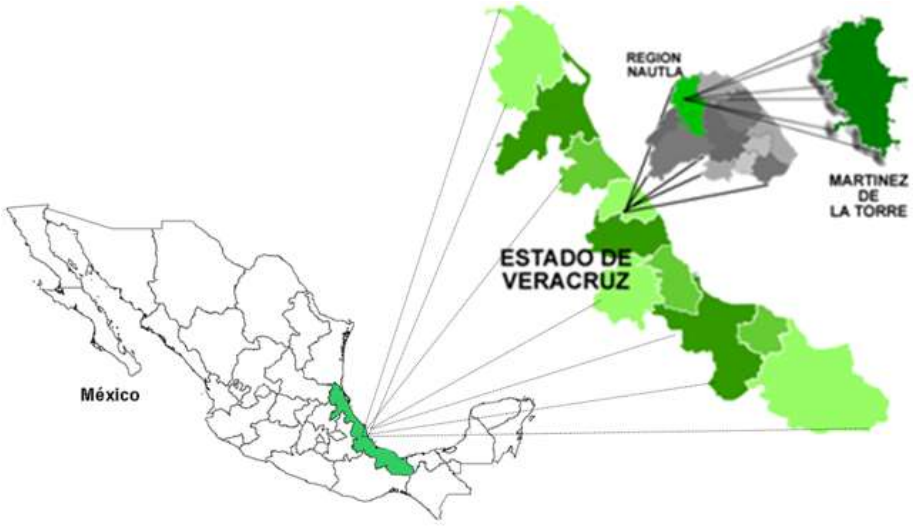


Figura 8. Localización del municipio de Martínez de la Torre, Veracruz.
Fuente: Elaborado con datos de INAFED (2018).

El municipio de Martínez de la Torre tiene una superficie de 402.48 km^2, que representa el 0.56% de la superficie total del estado. Así mismo, tiene una población de 108,842 habitantes, que corresponde al 1.35% del estado (INEGI, 2020). La densidad de la población es de 270 habitantes por km^2, que se encuentra por encima de la media estatal (112 habitantes por km^2) y de la media nacional (64 habitantes por km^2), reflejando mayor densidad de población en el municipio de Martínez de la Torre, Veracruz (INEGI, 2020).

La actividad económica más importante del sector primario en el municipio de Martínez de la Torre, estado de Veracruz, es la agricultura, que es la fuente principal de ingresos, aportando un valor de la producción de alrededor de 2,748.5 millones de pesos, que representa el 98% del valor de la producción agropecuaria del municipio (SIAP, 2021). Los frutales son los cultivos con mayor peso en el municipio de Martínez de la Torre, ya que ocupan el 96.6% de la superficie agrícola cultivada. En el grupo de los frutales, los cítricos representan el 99.7% de la superficie sembrada, de los cuales el limón ocupa el 56.4%, seguido por la naranja con el 35.3%, la toronja con el 8.3% y la mandarina con el 0.1% (SIAP, 2021).

Así mismo, otros de los cultivos importantes en el municipio de Martínez de la Torre, del grupo de los cultivos industriales, es la caña de azúcar, que representa el 1.7% de la superficie cultivada total, seguido por el maíz, del grupo de los cereales, con el 1.5% de la superficie sembrada total. La otra actividad importante en el municipio de Martínez de la Torre, estado de Veracruz, es la ganadería, que aporta 56.4 millones de pesos, 2.0% del valor de la producción agropecuaria del municipio. Las principales especies ganaderas son los bovinos, porcinos, ovinos, aves y abejas (SIAP, 2021).

Información de campo y sistematización

La información se obtuvo a través de encuestas aplicadas a productores en campo. Para obtener la muestra representativa en la región, se tomó en cuenta a los 2,106 productores en las 69 comunidades productoras de limón persa registrados en el padrón de citricultores pertenecientes al municipio de Martínez de la Torre, Veracruz (CONCITVER, 2012). De este total de comunidades se seleccionaron cinco, en las cuales se entrevistaron a 49 productores, los poblados y productores fueron elegidos por medio del muestreo aleatorio simple.

La información obtenida se organizó en matrices con las columnas de variable, cantidad e indicador. Los parámetros que se estimaron son valores mínimos, promedios, máximos y totales. Los indicadores que se calcularon son los coeficientes de participación.

Conceptos y variables de estudio

Superficie sembrada. La superficie sembrada es el área agrícola en la cual se deposita la semilla o plántula de cualquier cultivo, previa preparación del suelo (SIAP, 2019).

Superficie cosechada. La superficie cosechada es el área cultivada de la que se obtuvo producción agrícola (SIAP, 2019).

Rendimiento. El rendimiento refleja la productividad obtenida de una unidad de superficie cosechada, con base en la producción generada por unidad de superficie (SIAP, 2019).

Volumen de producción. El volumen de producción es la cantidad de producto que se obtiene en determinada cantidad de superficie cosechada (SIAP, 2019).

Productores. Persona física o moral que tiene habitualmente y como principal actividad económica la explotación agrícola de la tierra (Cruz, 2017).

Tenencia de la tierra. La tenencia de la tierra es la relación, jurídica o consuetudinaria, que existe entre las personas y la tierra (FAO, 2003). En el campo mexicano la tenencia de la tierra está conformada por la pequeña propiedad, los ejidos y las comunidades agrarias, designándose a estas dos últimas formas de tenencia como propiedad social o núcleos agrarios (Morett y Cosío, 2017).

Pequeña propiedad. La pequeña propiedad se refiere a la extensión de tierra que puede poseer un solo propietario, regulado por las leyes correspondientes. La pequeña propiedad puede ser agrícola, ganadera y forestal (Ley Agraria, 1992).

Ejido. El ejido, formado por los núcleos de población ejidal, tienen personalidad jurídica y patrimonio propio y son propietarios de las tierras que les han sido dotadas por el Estado y de las que hubieren adquirido por cualquier otro título (Ley Agraria, 1992 y Ruiz, 1990).

Comunidad agraria. La comunidad agraria, formada por núcleos de población comunal, tienen personalidad jurídica y patrimonio propio y son propietarios de las tierras restituidas y lo confirmado por el Estado y de las que hubieren adquirido por cualquier otro título (Ley Agraria, 1992 y Ruiz, 1990).

Indicador calculado

La información obtenida de las variables agrarias y agrícolas se utilizaron para el cálculo de valores mínimos, medios y máximos y coeficientes de participación.

Coeficiente de participación. Es el valor que representa la participación de un valor parcial con respecto de un total en términos de unidades (Caamal *et al.*, 2016). El procedimiento de cálculo es:

$$CP = (VP/ VT) \quad (1)$$

Donde: CP= Coeficiente de participación; VP= Valor parcial; VT= Valor total

El coeficiente de participación se transforma en participación porcentual al multiplicarse por cien, el procedimiento de cálculo es:

$$\% = (VP/ VT) * 100 \quad (2)$$

Donde: %= Participación porcentual; VP= Valor parcial; VT= Valor total

3. RESULTADOS Y DISCUSIÓN

Superficie cosechada y producción por municipios

Los principales municipios productores de limón persa del Distrito de Desarrollo Rural (DDR) de Martínez de la Torre, Veracruz, son Martínez de la Torre con 15,544 ha (44.4%) de la superficie cosechada y 272,931 ton (44.2%) de la producción; Atzalan con 5,668 ha (16.2%) de la superficie cosechada y 104,858 ton (17.0%) de la producción; San Rafael con 4,550 ha (13.0%) de la superficie cosechada y 78,735 ton (12.7%) de la producción; Tlapacoyan con 3,753 ha (10.7%) de la superficie cosechada y 74,440 ton (12.1%) de la producción; Papantla con 1,875 ha (5.4%) de la superficie cosechada y 27,188 ton (4.4%) de la producción;

entre otros. Los cinco municipios productores mencionados, aportaron cerca del 90% de la superficie cosechada y alrededor del 90% de la producción total de limón persa del DDR de Martínez de la Torre, Veracruz (SIACON, 2021 y Figuras 9 y 10).

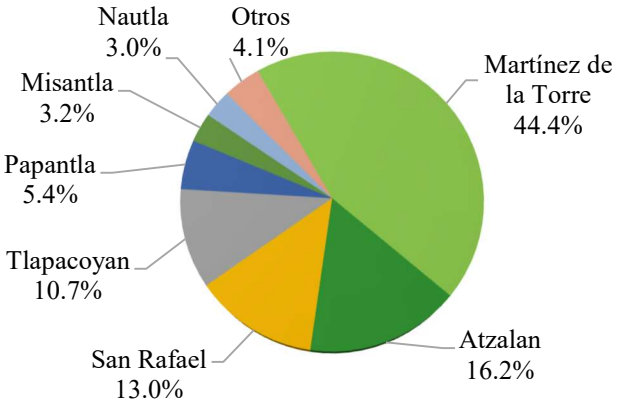


Figura 9. Distribución de la superficie cosechada de limón persa por municipios del DDR de Martínez de la Torre, Veracruz, 2020.
Fuente: Elaborado con datos de SIACON, 2021.

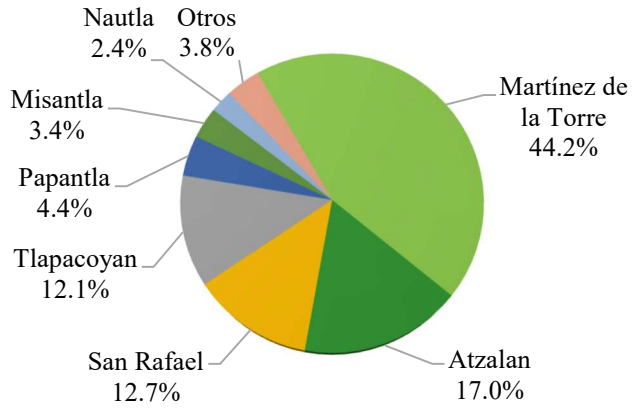


Figura 10. Distribución de la producción de limón persa por municipios del DDR de Martínez de la Torre, Veracruz, 2020.
Fuente: Elaborado con datos de SIACON, 2021.

Comunidades productoras de limón persa

En el municipio de Martínez de la Torre, Veracruz, se tienen registradas 69 comunidades productoras de limón persa en el Consejo Estatal Citrícola A.C. de Veracruz (CONCITVER), con 2,106 productores que reportan 8,378.9 hectáreas en producción, sin embargo, no todos los productores se encuentran registrados en el Consejo Estatal Citrícola (CONCITVER, 2012 y Tabla 1).

Las principales comunidades productoras de limón persa en el municipio de Martínez de la Torre, Veracruz son la localidad de Manantiales con 553.6 ha (6.6%), Puntilla Aldama con 503.4 ha (6.0%), Arroyo Blanco con 396.6 ha (4.7%), Hidalgo con 383.0 ha (4.6%), Arroyo del Potrero con 378.6 ha (4.5%), Paso Largo con 368.1 ha (4.4%), Flamencos con 363.0 ha (4.3%), entre otros, de un total de 8,381.9 hectáreas (CONCITVER, 2012 y Tabla 1).

Tabla 1. Comunidades con mayor superficie en producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.

Localidad	Superficie (ha)	%
Manantiales	553.6	6.6
Puntilla Aldama	503.4	6.0
Arroyo Blanco	396.6	4.7
Hidalgo	383.0	4.6
Arroyo del Potrero	378.6	4.5
Paso Largo	368.1	4.4
Flamencos	363.0	4.3
Manuel Ávila Camacho	261.4	3.1
La Palma	257.7	3.1
El Cañizo	236.6	2.8
Otros	4,679.9	55.8
Total	8,381.9	100.0

Fuente: Elaborado con datos del CONCITVER, 2012.

Edad y nivel de educación de los productores

Edad. El rango de edad de los productores de limón persa en el municipio de Martínez de la Torre, estado de Veracruz, se encuentra entre los 25 y 85 años, aunque la mayoría se encuentran entre los 41 y 55 años (Figura 11), mientras que la edad promedio es de 50 años, esta característica indica que la mayoría de los productores en esa región son grandes en edad, que puede ser una limitante para la innovación tecnológica, ya que los productores grandes de edad generalmente son más reacios a los cambios tecnológicos. Los procesos de adopción de nuevas tecnologías pueden facilitarse en la medida en que los productores jóvenes tengan una mayor participación en la producción de limón persa.

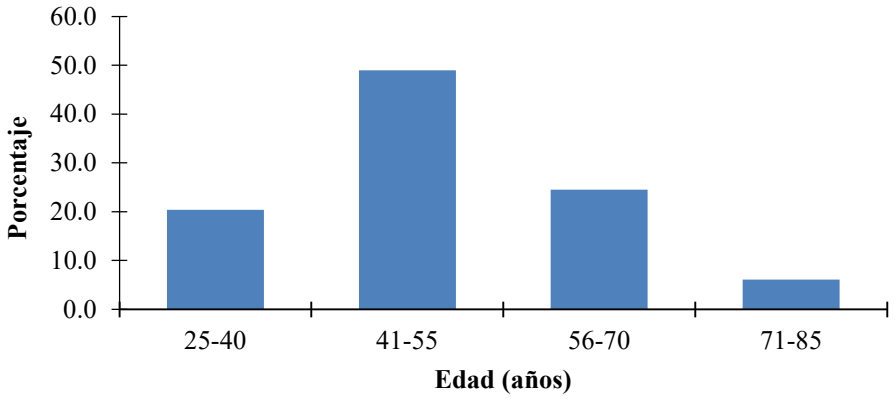


Figura 11. Edad de los productores de limón persa en el municipio de Martínez de la Torre, estado de Veracruz. Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Nivel de educación. El grado de escolaridad de los productores de limón persa en el municipio de Martínez de la Torre, estado de Veracruz, se concentra en el nivel básico (primaria y secundaria), le sigue en importancia el grupo de productores con nivel superior (licenciatura y posgrado), continuando con los productores con nivel medio superior (preparatoria) y, finalmente, aquellos que no alcanzaron a concluir el nivel básico de estudio (Figura 12). La información refleja que alrededor del 50% de los productores de limón persa tienen un bajo nivel educativo (ninguno y primaria).

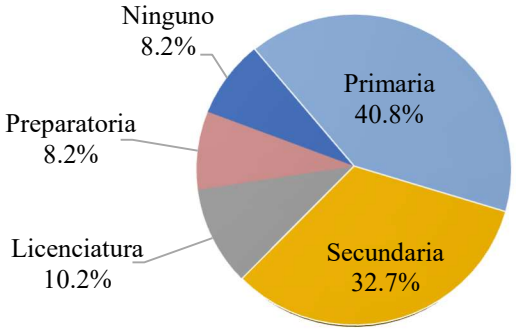


Figura 12. Niveles de educación de los productores de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Tamaño y tipo de tenencia de las unidades de producción

Tamaño de la superficie en producción. En la zona productora de limón persa en el municipio de Martínez de la Torre, estado de Veracruz, la superficie promedio por productor es pequeña, de 5.3 hectáreas (Tabla 3). Los productores, por el número de hectáreas en producción, el 46.9% tiene menos de 3.0 hectáreas, el 42.8% tiene entre 3.1 y 9.0 hectáreas y el 10.2% tiene más de 9.1 hectáreas en producción (Tabla 2 y Figura 13), lo que refleja que la mayoría de los productores son pequeños. El ingreso familiar de los productores, depende principalmente de la venta de limón persa, no obstante, muchos productores producen otros productos como la naranja, la mandarina, la toronja, el maíz y el frijol; y, además, también se ocupan en otras actividades económicas, diversificando así sus ingresos.

Tabla 2. Distribución de la superficie en producción del limón persa en el municipio de Martínez de la Torre, Veracruz.

Superficie en producción (ha)	Número de productores (%)		Superficie (%)
≤ 3	23	46.9	16.1
3.1 – 6	15	30.6	29.3
6.1 – 9	6	12.2	19.8
9.1 – 29	5	10.2	34.9
Total	49	100.0	100.0

Fuente: Elaborado con información de campo en 2013.

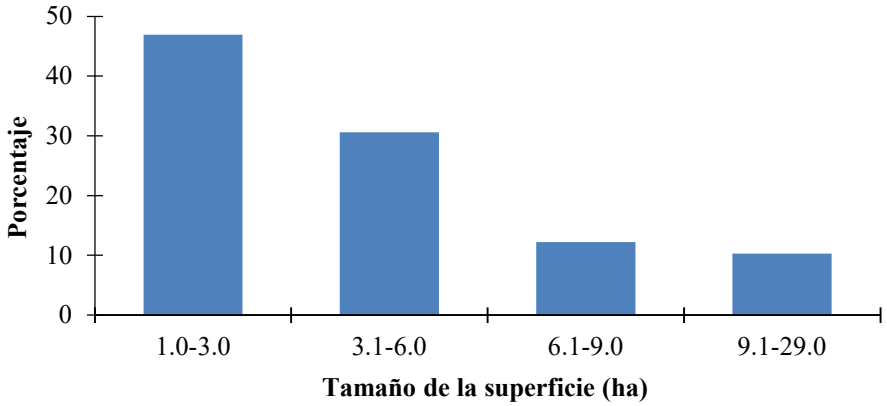


Figura 13. Distribución de los productores de limón persa por tamaño de la superficie en producción en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Tenencia de la tierra. La mayor parte de los productores de limón persa del municipio de Martínez de la Torre, estado de Veracruz, cuentan con tierra bajo el régimen ejidal, solamente una proporción muy pequeña de los productores se encuentra en el régimen de propiedad privada, prestada y/o rentada (Figura 14). Una de las ventajas del régimen ejidal es que les permite a los productores acceder a los apoyos gubernamentales que se ofrecen a los pequeños productores, especialmente a los ejidales.

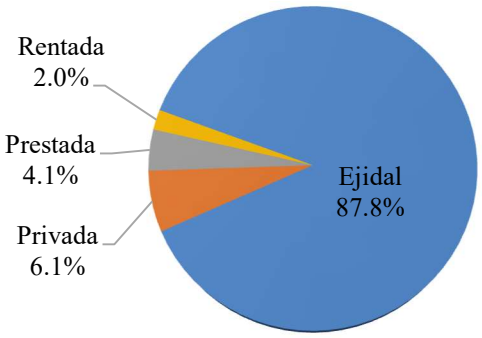


Figura 14. Régimen de propiedad de los productores de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Organización económica de los productores. En el municipio de Martínez de la Torre, Veracruz, han existido varias organizaciones económicas de productores, las cuales, con el paso del tiempo han venido desapareciendo por problemas internos y escasos beneficios económicos obtenidos, lo que se ve reflejado en el escaso interés de los productores por pertenecer a alguna de las formas de organización legalmente reconocidas (Figura 15), mientras que los pocos productores que pertenecen a alguna organización casi no perciben beneficios de la organización, lo que explica la baja participación de los productores de limón persa en organizaciones económicas de productores.

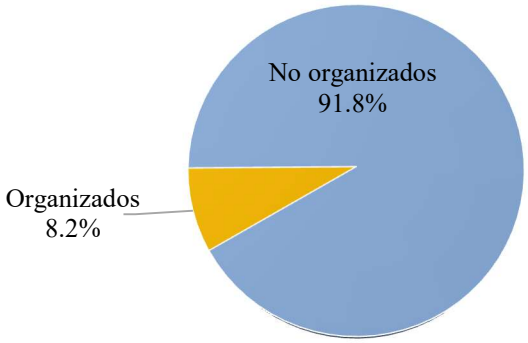


Figura 15. Organización económica de los productores de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Superficie cosechada y producción

Edad de las plantaciones. La mayoría de los productores de limón persa en el municipio de Martínez de la Torre, estado de Veracruz, tienen plantaciones de 4 a 10 años de edad (72.0%), con edad promedio de 8.5 años (Figura 16), lo anterior refleja que la mayoría de las plantaciones se encuentran en una etapa creciente de producción, lo que significa que pueden mejorar los rendimientos actuales.

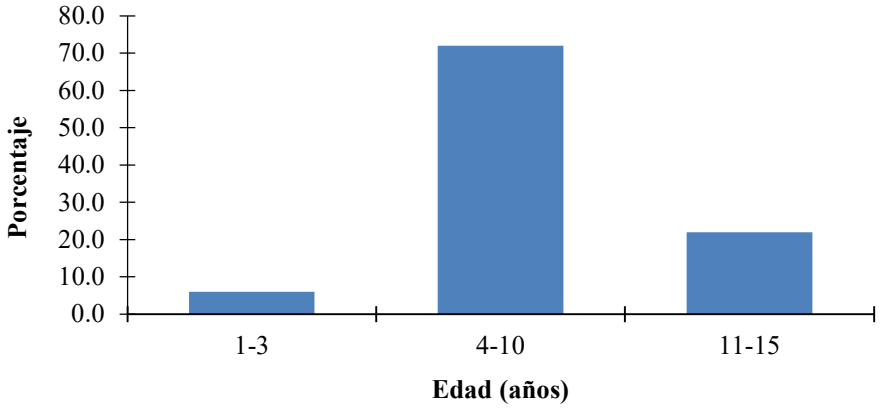


Figura 16. Distribución de la edad de las plantaciones en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Superficie y rendimiento. La superficie media en producción de limón persa es de 5.3 hectáreas por productor, mientras que la superficie máxima es de 29 hectáreas y la mínima es de una hectárea, y el rendimiento promedio es de 12.0 toneladas por hectárea, mientras que el rendimiento máximo es de 19.0 y el mínimo de 6.0 toneladas por hectárea (Tabla 3 y Figura 17).

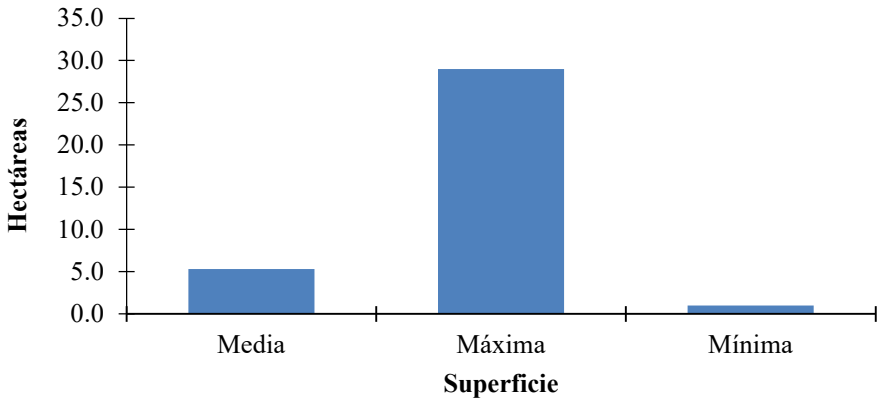


Figura 17. Superficie en producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

El rendimiento promedio de 12 toneladas por hectárea de los pequeños productores se encuentra por debajo de la media regional y estatal, 17.6 y 16.7 toneladas por hectárea (Tabla 3 y Figura 18), lo cual indica que existen problemas de productividad de este estrato de productores en la región.

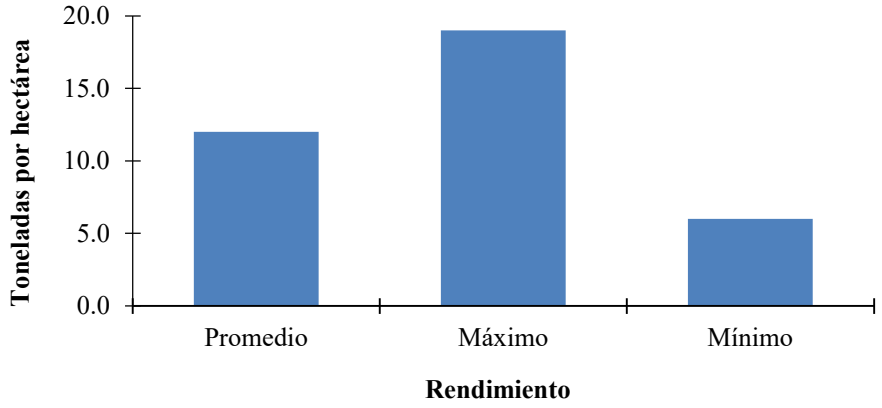


Figura 18. Rendimiento del limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Producción. La producción media obtenida por productor de limón persa es de 63.6 toneladas, mientras que la producción máxima es de 348.0 y la mínima es de 12.0 toneladas (Tabla 3 y Figura 19).

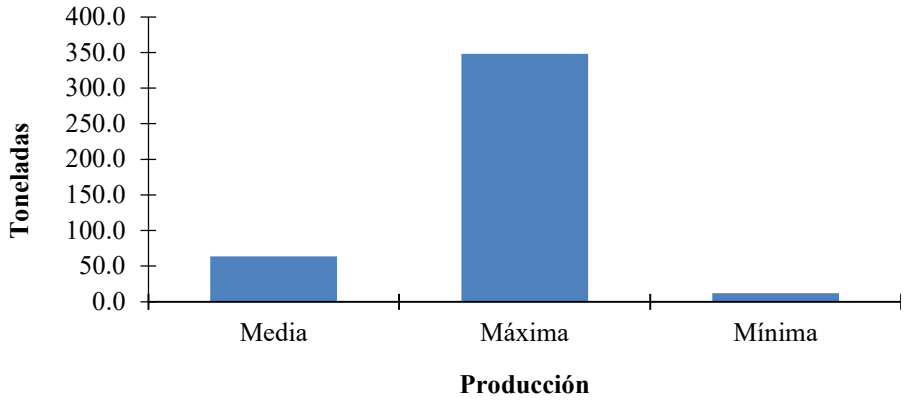


Figura 19. Producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Tabla 3. Datos generales de la producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz, 2013.

Concepto	Cantidad
Superficie media (ha)	5.3
Superficie máxima (ha)	29.0
Superficie mínima (ha)	1.0
Rendimiento promedio (ton/ha)	12.0
Rendimiento máximo (ton/ha)	19.0
Rendimiento mínimo (ton/ha)	6.0
Producción media (ton)	63.6
Producción máxima (ton)	348.0
Producción mínima (ton)	12.0

Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Fertilización y control de malezas

Fertilizantes. El uso de fertilizantes granulados en el municipio de Martínez de la Torre, estado de Veracruz, es del 98% de la población de productores de limón persa (Tabla 4 y Figura 20), la proporción de productores que no aplica fertilizantes es porque están abandonando la producción de limón persa.

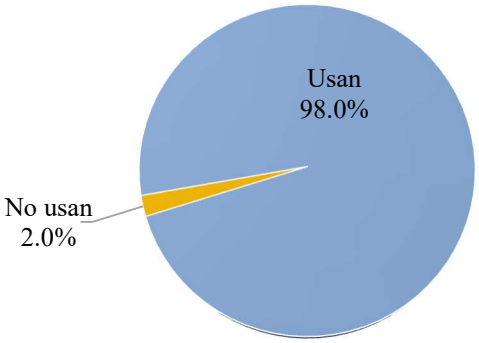


Figura 20. Uso de fertilizantes granulados en la producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Por otro lado, en el caso de los fertilizantes foliares, el 76% de los productores aplica este tipo de fertilizante (Tabla 4 y Figura 21). Las dosis aplicadas difieren de un productor a otro, en general se observa que aplican una cantidad menor a la recomendada en los paquetes tecnológicos, lo que explica en parte los bajos rendimientos del cultivo.

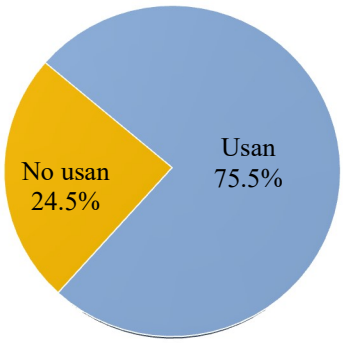


Figura 21. Uso de fertilizantes foliares en la producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Herbicidas. El 94% de los productores entrevistados aplican algún tipo de herbicida (Tabla 4 y Figura 22) y el 71% aplica pesticidas para el control de las enfermedades y plagas (Tabla 4 y Figura 23), los niveles de aplicación señalados reflejan limitaciones tecnológicas de manejo del cultivo de limón persa.

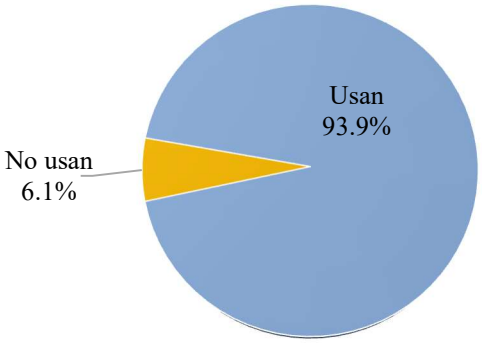


Figura 22. Uso de herbicidas en la producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

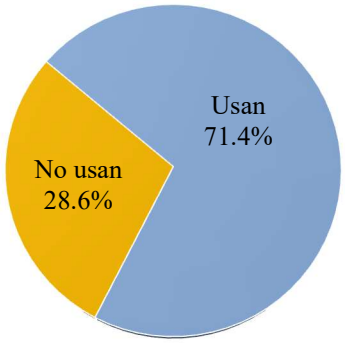


Figura 23. Uso de pesticidas en la producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz.
Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

Tabla 4. Uso de agroquímicos en la producción de limón persa en el municipio de Martínez de la Torre, estado de Veracruz, 2013.

Insumo	Productores			
	Usan	%	No usan	%
Fertilizantes Granulados	48	98	1	2
Fertilizantes Foliares	37	76	12	24
Herbicidas	46	94	3	6
Pesticidas	35	71	14	29

Fuente: Elaborado con datos de las encuestas aplicadas, 2013.

4. CONCLUSIÓN

La producción de limón persa se realiza en superficies ejidales pequeñas, por productores de bajos niveles de educación y organización y de edad adulta y avanzada, con bajos rendimientos y niveles de producción y niveles medios de uso de fertilizantes y herbicidas. El uso de agroquímicos como los fertilizantes granulados, fertilizantes foliares, herbicidas, plaguicidas y pesticidas son de uso moderado, las cantidades y tipo de producto dependen, en primer lugar, de la capacidad económica de los productores para adquirir los insumos en el mercado y en segundo lugar a los requerimientos de las plantaciones.

REFERENCIAS

[1] ALMAGUER V., G., AYALA G., A. V. (2014). Adopción de innovaciones en limón 'persa' (Citrus latifolia Tan.) en Tlapacoyan, Veracruz. Uso de bitácora. *Revista Chapingo Serie Horticultura*, 20(1): 89-100. doi: 10.5154/r.rchsh.2010.10.076

[2] APOYOS Y SERVICIOS A LA COMERCIALIZACIÓN AGROPECUARIA (ASERCA). (1995): *Limón persa. Estudio del mercado mundial*. México: Technomanagement S. A., 1995.

[3] APOYOS Y SERVICIOS A LA COMERCIALIZACIÓN AGROPECUARIA (ASERCA). (1996): El limón persa y el limón mexicano: la complementariedad del mercado. *Claridades Agropecuarias*, 30. México. [En línea] Disponible en web: <https://info.aserca.gob.mx/claridades/marcos.asp?numero=30>

[4] CAAMAL C., I., JERÓNIMO A., F., PAT F., V.G. (2016): Análisis de las variables económicas de la producción de naranja en México. *Revista Mexicana de Ciencia Agrícolas*. México: INIFAP.

[5] CONSEJO ESTATAL CITRÍCOLA DE VERACRUZ A. C. (CONCITVER). (2012): *Portal de internet*. [En línea] Disponible en web: <http://www.concitver.com> Consultado 6, 7, 2012.

[6] CONSTITUCIÓN POLÍTICA DE LOS ESTADOS UNIDOS MEXICANOS (CPEUM). (1917): Artículo 27. Última reforma publicada DOF 28-05-2021. Cámara de Diputados del H. Congreso de la Unión. México. Disponible en web: <http://www.diputados.gob.mx/LeyesBiblio/ref/cpeum.htm> Consultado 24, 8, 2021.

[7] CORPORACIÓN COLOMBIA INTERNACIONAL (CCI). (2000). Inteligencia de mercados: limas y limones. Perfil de producto. No. 18. Ministerio de Agricultura y Desarrollo Rural, Bogotá, Colombia. [En línea]. Disponible en web: <https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=1&cad=rja&uact=8&ved=2ahUKEwjsoNDB7bneAhUQDKwKHcxJCRcQFjAAegQIABAC&url=http%3A%2F%2Fbibliotecadigital.agronet.gov.co%2Fbitstream%2F11348%2F5346%2F1%2F2005113153827_perfil_Limas_Limones.pdf&usg=AOvVaw3tlDswd5z23AHOmZBIKOor>

[8] CRUZ, G. (2017): *Concepto de productor agrícola en derecho agrario*. [En línea]. Disponible en web: <https://mexico.leyderecho.org/productor-agricola/>, Consultado 24,8, 2021.

[9] FAOSTAT. (2021). Datos. FAO, Bases de datos estadísticos [En línea] Disponible en: <http://www.fao.org/faostat/es/#data> Consultado 6, 7, 2021.

[10] FLORES, V., CRUZ, M., ROSANO, G. y RODRÍGUEZ, E. (2014). Medición de la Competitividad de los productores de limón Persa por medio de Logit, caso de los productores de Cuitláhuac Veracruz. *En* RAMOS., M y AGUILERA, V. (eds.), Ciencias de la Ingeniería y Tecnología, Aguilera. Handbook. Guanajuato, México: ©ECORFAN.

[11] IICA, MAG, FRUTALES y PROESA. (2004): *Fruticultura: oportunidades de inversión en El Salvador.* [En línea]. Disponible en: <http://opackoha.iica.int/cgi-bin/koha/opac-detail.pl?biblionumber=1004>

[12] INSTITUTO NACIONAL DE ESTADÍSTICA Y GEOGRAFÍA (INEGI). (2020): Censo de Población y Vivienda 2020. México. [En línea]. Disponible en: <https://www.inegi.org.mx/programas/ccpv/2020/default.html#Tabulados>, Consultado 20, 8, 2021.

[13] INSTITUTO NACIONAL PARA EL FEDERALISMO Y EL DESARROLLO MUNICIPAl (INAFED). (2018): *Estado de Veracruz de Ignacio de la Llave. Enciclopedia de los municipios y delegaciones de México*. México: SEGOB, [En línea]. Disponible en web: <http://www.inafed.gob.mx/work/enciclopedia/EMM30veracruz/index.html>

[14] KHAN, M. M., AL-YAHYAI, R., AL-SAID, F. (2017): Introduction and overview of lime. En *The Lime. Botany, production and uses* (Khan *et al.* eds.). Boston MA: CAB international, 2017.

[15] LEY AGRARIA (1992): Artículos 9o, 12, 14, 98, 99 y 119. Última reforma publicada DOF 25-06-2018. Cámara de Diputados del H. Congreso de la Unión. México. Disponible en web: <http://www.diputados.gob.mx/LeyesBiblio/index.htm> Consultado 24, 8, 2021.

[16] MONTERO C., Y. (2009): Evaluación de la actividad antimicrobiana del aceite esencial de limón persa (Citrus latifolia Tanaka). Tesis de maestría. Universidad Veracruzana, México.

[17] MORETT S., J.C. y COSÍO R., C. (2017): Panorama de los ejidos y comunidades agrarias en México. *Agricultura, Sociedad y Desarrollo*, 14(1), 125-152.

[18] MORTON, J. (1987): Lemon. En: *Fruits of warm climates*. Julia F. Morton, Miami, FL.

[19] ORGANIZACIÓN DE LAS NACIONES UNIDAS PARA LA AGRICULTURA Y LA ALIMENTACIÓN (FAO). (2003): *Tenencia de la tierra y desarrollo rural*. FAO Estudios sobre tenencia de la tierra 3. Roma: FAO,

[20] RUIZ M. M. (1990): *Derecho Agrario*. México D.F.: Instituto de Investigaciones Jurídicas-UNAM, 1990.

[21] SERVICIO DE INFORMACIÓN AGROALIMENTARIA Y PESQUERA (SIAP). (2019): *Normatividad técnica para la generación de estadística básica agropecuaria 2019*. SIAP-SADER. México. [En línea] disponible en web: <https://www.gob.mx/siap/documentos/normatividad-estadistica> Consultado 19, 10, 2020.

[22] SERVICIO DE INFORMACIÓN AGROALIMENTARIA Y PESQUERA (SIAP). (2021): *Información*. SADER, México. [En línea] Disponible en web: <https://www.gob.mx/siap#1410>

[23] SISTEMA DE INFORMACIÓN AGROALIMENTARIA DE Consulta (SIACON). (2021): Producción agrícola estatal. SIAP-SADER. México. Disponible en web: <https://www.gob.mx/siap/documentos/siacon-ng-161430>

[24] UNITED STATES DEPARTMENT OF AGRICULTURE (USDA). (2019): *Limes, raw*. FoodData Central. Agricultural Research Service. [En línea] 2021 Disponible en web: <https://fdc.nal.usda.gov/fdc-app.html#/food-details/168155/nutrients>, Consultado 3, 9, 2021.

Anexo

Tabla 1. Principales cultivos cítricos en el mundo, 2019.

Cultivo	Superficie cosechada (ha)	%	Rendimiento (ton/ha)	Producción (ton)	%
Naranjas	4,060,129	41.0	19.4	78,699,604	49.8
Tangerinas, mandarinas, otros	2,756,887	27.9	12.9	35,444,080	22.4
Limones y limas	1,226,617	12.4	16.3	20,049,630	12.7
Toronja y pomelo	346,191	3.5	26.8	9,289,462	5.9
Otros	1,508,639	15.2	9.6	14,496,484	9.2
Total	9,898,463	100.0	17.0	157,979,260	100.0

Fuente: Elaborado con datos de FAOSTAT, 2021.

Tabla 2. Principales países productores de limones y limas a nivel mundial, 2019.

País	Superficie cosechada (ha)	%	Rendimiento (ton/ha)	Producción (ton)	%
India	305,000	24.9	11.4	3,482,000	17.4
México	183,575	15.0	14.7	2,701,828	13.5
China, Continental	128,823	10.5	20.7	2,666,082	13.3
Argentina	57,541	4.7	33.1	1,904,765	9.5
Brasil	56,491	4.6	26.8	1,511,185	7.5
Turquía	40,155	3.3	23.7	950,000	4.7
España	46,840	3.8	18.9	884,890	4.4
EE.UU.	21,970	1.8	39.9	876,340	4.4
Sudáfrica	24,542	2.0	20.8	510,524	2.5
Irán	29,478	2.4	16.0	471,008	2.3
Otros	332,202	27.1	12.0	4,091,008	20.4
Total	1,226,617	100.0	13.0	20,049,630	100.0

Fuente: Elaborado con datos de FAOSTAT, 2021.

Tabla 3. Principales frutales en México, 2020.

Cultivo	Superficie sembrada (ha)	%	Superficie cosechada (ha)	%	Rendimiento (ton/ha)	Producción (ton)	%
Naranja	343,245	20.0	327,756	20.7	14.2	4,648,620	19.5
Aguacate	241,136	14.1	224,422	14.2	10.7	2,393,849	10.0
Limón	207,838	12.1	185,117	11.7	15.4	2,851,427	11.9
Mango	204,643	12.0	193,246	12.2	10.8	2,085,751	8.7
Nuez	144,653	8.4	108,771	6.9	1.5	164,652	0.7
Plátano	80,547	4.7	79,747	5.0	30.9	2,464,171	10.3
Manzana	60,013	3.5	56,706	3.6	12.6	714,203	3.0
Tuna	45,140	2.6	43,393	2.7	10.9	471,427	2.0
Sandía	39,735	2.3	39,118	2.5	34.8	1,362,393	5.7
Durazno	33,528	2.0	26,091	1.6	6.6	172,950	0.7
Otros	311,628	18.2	297,588	18.8	12.4	6,553,372	27.4
Total	1,712,107	100.0	1,581,955	100.0	12.8	23,882,815	100.0

Fuente: Elaborado con datos de SIACON, 2021.

Tabla 4. Principales variedades de limones y limas a en México, 2020.

Cultivo	Superficie sembrada (ha)	%	Superficie cosechada (ha)	%	Rendimiento (ton/ha)	Producción (ton)	%
Limón persa	100,449	47.8	94,595	50.5	15.3	1,447,416	50.3
Limón agrio (mexicano)	95,594	45.5	81,728	43.6	15.0	1,221,459	42.4
Limón italiano	9,129	4.3	7,142	3.8	19.7	140,748	4.9
Lima	2,394	1.1	2,364	1.3	9.1	27,596	1.0
Otros	2,666	1.3	1,651	0.9	21.3	41,804	1.5
Total	210,232	100.0	187,481	100.0	18.0	2,879,023	100.0

Fuente: Elaborado con datos de SIACON, 2021.

Tabla 5. Distribución de la producción de limón persa en México, 2020.

Estado	Superficie sembrada (ha)	%	Superficie cosechada (ha)	%	Rendimiento (ton/ha)	Producción (ton)	%
Veracruz	47,809	47.6	46,959	49.6	16.7	785,668	54.3
Oaxaca	15,229	15.2	14,350	15.2	13.8	198,586	13.7
Tabasco	7,227	7.2	7,223	7.6	11.9	86,097	5.9
Jalisco	5,960	5.9	4,996	5.3	18.6	92,973	6.4
Yucatán	4,499	4.5	3,683	3.9	21.0	77,205	5.3
Michoacán	3,201	3.2	2,852	3.0	10.9	31,165	2.2
Puebla	2,829	2.8	2,682	2.8	13.4	35,969	2.5
Nayarit	2,813	2.8	2,464	2.6	10.8	26,518	1.8
Chiapas	2,735	2.7	2,412	2.6	6.4	15,329	1.1
Campeche	2,078	2.1	1,925	2.0	9.1	17,518	1.2
Otros	6,069	6.0	5,051	5.3	13.8	80,387	5.6
Total	100,449	100.0	94,595	100.0	13.6	1,447,416	100.0

Fuente: Elaborado con datos de SIACON, 2021.

Tabla 6. Superficie total, de temporal y de riego de limón persa en México, 2020 (ha).

Estado	Total	%	Temporal	%	Riego	%
Veracruz	47,809	100.0	39,075.50	81.7	8,733.81	18.3
Oaxaca	15,229	100.0	14,138.50	92.8	1,090.00	7.2
Tabasco	7,227	100.0	5,724.32	79.2	1,503.00	20.8
Jalisco	5,960	100.0	31	0.5	5,929.40	99.5
Yucatán	4,499	100.0	64.45	1.4	4,434.98	98.6
Michoacán	3,201	100.0	2	0.1	3,198.50	99.9
Puebla	2,829	100.0	2,740.70	96.9	88.5	3.1
Nayarit	2,813	100.0	1,689.00	60.1	1,123.50	39.9
Chiapas	2,735	100.0	2,568.95	93.9	166	6.1
Campeche	2,078	100.0	706.2	34.0	1,372.00	66.0
Otros	6,069	100.0	1,365	22.5	4,704	77.5
Total	100,449	100.0	68,105.62	67.8	32,343.19	32.2

Fuente: Elaborado con datos de SIACON, 2021.

Tabla 7. Distribución de la producción de limón persa en el estado de Veracruz, por Distrito de Desarrollo Rural, 2020.

Distrito	Superficie sembrada (ha)	%	Superficie cosechada (ha)	%	Rendimiento (ton/ha)	Producción (ton)	%
Martínez de la Torre	35,438	74.1	35,010	74.6	17.6	617,687	78.6
Veracruz	3,570	7.5	3,570	7.6	17.3	61,634	7.8
Fortín	3,342	7.0	3,342	7.1	16.3	54,469	6.9
Coatepec	1,614	3.4	1,519	3.2	9.9	15,003	1.9
Ciudad Alemán	1,355	2.8	1,325	2.8	12.5	16,496	2.1
Tuxpan	890	1.9	777	1.7	9.9	7,660	1.0
Otros	1,600	3.3	1,416	3.0	9.9	12,718	1.6
Total	47,809	100.0	46,959	100.0	16.7	785,668	100.0

Fuente: Elaborado con datos de SIACON, 2021.

Tabla 8. Principales municipios productores del DDR de Martínez de la Torre, Veracruz.

Municipio	Superficie sembrada (ha)	%	Superficie cosechada (ha)	%	Rend. (ton/ha)	Producción (ton)	%
Martínez de la Torre	15,584	44.0	15,544	44.4	17.6	272,931	44.2
Atzalan	5,768	16.3	5,668	16.2	18.5	104,858	17.0
San Rafael	4,557	12.9	4,550	13.0	17.3	78,735	12.7
Tlapacoyan	3,760	10.6	3,753	10.7	19.8	74,440	12.1
Papantla	1,878	5.3	1,875	5.4	14.5	27,188	4.4
Nautla	1,279	3.6	1,064	3.0	14.2	15,109	2.4
Misantla	1,124	3.2	1,111	3.2	19.0	21,093	3.4
Otros	1,489	4.2	1,446	4.1	16.3	23,334	3.8
Total	35,438	100.0	35,010	100.0	17.6	617,687	100.0

Fuente: Elaborado con datos de SIACON, 2021.

Capítulo 8 pp 121-134

DIAGNÓSTICO DE DEPRESIÓN EN ADULTOS MAYORES EN EL NIVEL PRIMARIO DE ATENCIÓN

Rosales-Ibárez, América A.[1]; Mendoza-Rodríguez, Cristina[2]; Ibárez-Castro, Aidé[1] y Rosales-Jiménez, Antonio[3]

[1]Secretaría de Salud Guerrero.
[2]Universidad de ciencias médicas de Villa Clara, Cuba, Facultad de medicina.
[3]Universidad Autónoma de Guerrero, Facultad de Medicina.

RESUMEN

Actualmente, en Cuba, la mayor parte de la población tiene una esperanza de vida igual o superior a los 72.7 años en hombres y 78.1 años en mujeres, por lo que el envejecimiento es un fenómeno manifiesto, y la población mayor de 60 años representa el 19.8% de la población total. Villa Clara, con un total de 244 331 habitantes, se reconoce entre las provincias más envejecidas del país, lo cual se evidencia en el 20% de su población, que es mayor de 85 años, como consecuencia las patologías propias de los adultos mayores, constituirán un problema de salud pública en Villa Clara y en todo el país. Uno de los trastornos más recurrentes, que supera el 50% de la población de adultos mayores, es la depresión. Aunque se han realizado diversos estudios en el mundo sobre ese fenómeno, en Cuba no existen cifras concluyentes sobre este padecimiento, aunque se identifican sus factores de riesgo. En el presente estudio se identifican los factores de riesgo de depresión en adultos mayores y generar un modelo matemático que aporte elementos para el diseño de estrategias de intervención comunitaria en la población de personas de la tercera edad del Policlínico "Santa Clara", del Hospital Provincial Psiquiátrico Universitario, "Dr. Luis San Juan Pérez". Se ha encontrado que los factores de riesgo de depresión para esta población son *pérdidas, inadaptación a la jubilación, eventos negativos en su vida, estilo de afrontamiento de pérdidas, red de apoyo familiar y vivir solo.* A partir de estas variables y usando un modelo de regresión logística se puede estimar la probabilidad de depresión en adultos mayores que acuden al policlínico, con lo que se abre la posibilidad de intervenir con medidas de prevención en el nivel primario de atención.

Palabras claves: depresión, escala de depresión, regresión logística binaria.

ABSTRACT

Currently, in Cuba, most of the population has a life expectancy equal to or greater than 72.7 years in men and 78.1 years in women, so aging is a manifest phenomenon, and the population over 60 years represents the 19.8% of the total population. Villa Clara, with a total of 244,331 inhabitants, is recognized as one of the oldest provinces in the country, which is evident in 20% of its population, which is over 85 years of age, as a consequence of the pathologies typical of the elderly, they will constitute a public health problem in Villa Clara and the whole country. One of the most recurring disorders, which exceeds 50% of the population of older adults, is depression. Although various studies have been carried out around the world on this phenomenon, in Cuba there are no conclusive figures on this condition, although its risk factors are identified. In the present study, the risk factors for depression in older adults are identified and a mathematical model is generated that

provides elements for the design of community intervention strategies in the elderly population of the "Santa Clara" Polyclinic, of the provincial hospital university psychiatrist, “Dr. Luis San Juan Perez. It has been found that the risk factors for depression for this population are loss, maladjustment to retirement, negative life events, loss coping style, family support network, and living alone. Based on these variables and using a logistic regression model, the probability of depression in older adults who attend the polyclinic can be estimated, thus opening up the possibility of intervening with prevention measures at the primary care level.

Keywords: depression, depression scale, binary logistic regression.

1. Introducción.

El envejecimiento se conceptualiza como un proceso degenerativo multiogánico de naturaleza multifactorial con grado de progresión e intensidad variable para cada individuo y de un órgano a otro en un mismo individuo como resultado de factores genéticos todavía mal definidos. Las consecuencias del envejecimiento pueden agravarse por la presencia de enfermedades crónicas concomitantes y consecuencias administrativas y sociales de la ancianidad (González-Menéndez, 2008).

En el presente trabajo de evitan las actitudes y normas negativas que a menudo influyen en la forma de concebir los desafíos derivados del envejecimiento de la población y las respuestas de la sociedad a ellos. En cambio, se parte de la hipótesis de que el envejecimiento es un proceso valioso, aunque frecuentemente complejo, y que las personas mayores hacen muchas contribuciones fundamentales a la sociedad. Se considera que es bueno envejecer y que las sociedades están mejor por tener a las personas mayores. Al mismo tiempo, se reconoce que muchas personas mayores sufren pérdidas significativas, ya sea en su capacidad física o cognitiva, o por la pérdida de familiares, de amigos y de roles que desempeñaban antes en la vida. Algunas de estas pérdidas pueden evitarse, y las personas y la sociedad deben trabajar para prevenirlas, pero otras son inevitables. Las respuestas de la sociedad al envejecimiento no deben negar estos problemas, sino tratar de fomentar la recuperación y la adaptación.

La Organización Mundial de la salud (OMS, 2002), promulga la visualización de un envejecimiento activo, concepto que acentúa la relación del sujeto con su entorno y consiste en entender al envejecimiento como el proceso en que se optimizan las oportunidades de salud, participación y seguridad, permitiendo que potencien sus capacidades físicas, y bienestar social, para de esta manera lograr una mejora en la calidad de vida. El informe plantea que el mismo depende de una diversidad de determinantes que rodean a las personas, sus familias, y la sociedad en general; dentro de los cuales podemos mencionar los determinantes económicos, determinantes sociales, determinantes personales, determinantes conductuales, el entorno físico, los servicios sociales y de salud también son determinantes.

Más allá de las pérdidas biológicas, la vejez con frecuencia conlleva otros cambios importantes. Se trata de cambios en los roles y las posiciones sociales, y la necesidad de hacer frente a la pérdida de relaciones estrechas (Tello-Rodríguez & Varela-Pinedo, 2016). Dentro de los múltiples cambios podemos mencionar la disminución de la fuerza muscular, afectación de la nitidez de los órganos de los sentidos, reducción de las potencialidades sexuales, repercusión estética de los cambios posturales, modificaciones de la marcha, alteraciones de la piel. Todo esto junto a: los déficit apreciados en la memoria, la afectividad, los intereses, la volición, las capacidades y los hábitos que determinan una situación de desventaja que de no ser compensada con la gratificaciones psicosociales que deben emanar del medio familiar, laboral y social provocan una actitud de notable sensibilidad, recelo y hostilidad que da al traste con la relaciones interpersonales y se expresa muchas veces por actitudes hipercríticas, tozudez, retraimiento afectivo, autosuficiencia y otras conductas orientadas o negar sus necesidades de afecto (Calderón M., 2018).

La depresión es una enfermedad frecuente en todo el mundo, y se calcula que afecta a más de 300 millones de personas. Es un trastorno mental, que se manifiesta según la OMS con la presencia de tristeza, pérdida de interés o placer, sentimientos de culpa o falta de autoestima, trastornos del sueño o del apetito, sensación de cansancio y disminución de la concentración. La depresión puede cronificarse y dificultar considerablemente el desempeño laboral, cognitivo y la capacidad de afrontar la vida diaria. En su forma más grave, puede conducir al suicidio (Luna-Reyes & Vilchez-Hernández, 2017).

Según la Organización Panamericana de Salud (OPS, 2013) estima que hasta 1 de cada 4 personas pasan por una depresión mayor en algún momento de su vida. Debido al gran impacto de este trastorno y con el objetivo de conocer los factores biológicos y ambientales que contribuyen a su origen; han surgido diversas hipótesis, tales como la de las monoaminas, la dopaminérgica, gabaérgica y la teoría de la diátesis-estrés, entre otras (Díaz-Villa & González-González, 2012), por otra parte (Wolff L, 2010), establecieron que, además del factor biológico, se debe poner énfasis en otros componentes asociados y posibles desencadenantes de este trastorno ya que la depresión es el resultado de interacciones complejas entre factores sociales, psicológicos y biológicos que influencian el desarrollo, tratamiento y pronóstico de la enfermedad.

1.1 Justificación del problema

El estudio realizado en 2020, en el área de salud perteneciente al Policlínico Universitario "Dr. Rudesindo Antonio García del Rijo", Sancti Spíritus, denominado, *Estados emocionales de adultos mayores en aislamiento social durante la COVID-19*, muestra, en el periodo estudiado, que el 61.0% correspondió al sexo femenino, el 57.0% pertenece al grupo de 70-79 años, la mayoría de los ancianos vivía acompañado con su pareja, un menor de edad o un discapacitado (64.0%), solo el 36.0% vivía efectivamente solo, el 65.0% no tenía vínculo laboral y el 89.0% presentaba patologías consideradas de riesgo para la COVID-19. Predominó un nivel de irritabilidad normal, tanto externa (68.0%) como interna (70.0%), un nivel leve de ansiedad (73.0%), un nivel leve de depresión (50.0%) y el 47.0% mostró alteración en los niveles de estrés.

En Cuba, el aislamiento ha provocado que las personas de 60 años o más hayan tenido que abandonar las actividades cotidianas y centrarse solo en las que se pueden realizar dentro del hogar. Dejar de trabajar, el que aún lo hacía, no asistir a los círculos de abuelos, a la Universidad de la tercera edad, a sus prácticas religiosas, no visitar a las amistades, a sus familiares, ni a los vecinos, no poder realizar compras de ninguna clase (Freitas C, 2016); todos estos factores podrían explicar los estados emocionales no satisfactorios identificados en los ancianos estudiados.

Se han realizado investigaciones en la población en general de distintas edades, incluyendo la de 60 años y más. En China, según se apunta en (Buiques C, 2015), se observó en la población general un 54.0% de impacto psicológico moderado a severo, un 29.0% de síntomas ansiosos y un 8.0% de estrés, todos entre moderados y severos; las personas mayores de 60 años presentaron un alto distrés

En España, Sandín, Valiente, García Escalera y Chorot (Vaughan, Corbin, & Goveas, 2015), al valorar el impacto psicológico de la pandemia en su población, encontraron síntomas (moderados a severos): estrés 32.0%, psicosomáticos 6.0%, problemas para dormir 36.3%, disfunción social en la actividad diaria 10.0% y depresivos 5% psicológico.

1.2 Formulación del problema y objetivos.

Los resultados de las distintas investigaciones internacionales sobre la depresión en el adulto mayor y su abordaje en la República de Cuba, permite orientar un estudio que posibilite hacer un *diagnóstico de depresión* para la población mayor de 60 años y más, por el crecimiento poblacional de este grupo de edad que convertirá a Cuba en el país más longevo en Latino América en el año 2020.

Esta patología es un gran reto como factor de riesgo de demencia, trastornos psiquiátricos, deterioro cognitivo, perdida de actividades de la vida diaria, dificultad para el cumplimiento de tareas, etc., y a la vez los factores de riesgo que la detonan, construyendo desde un síndrome hasta una enfermedad psiquiátrica. El crecimiento de su prevalencia traerá desafíos en su manejo, tanto en el primer nivel de atención como en el segundo nivel (hospital psiquiátrico) prevenirla es el desafío fundamental y movilizar desde la familia, la comunidad y la sociedad como pilares de control y erradicación de enfermedades en el Sistema de Salud Cubano.

Por lo antes planteado, llegamos a una gran interrogante: ¿Qué relación existe entre factores de riesgo y la Depresión en adultos mayores en el nivel primario de atención en la ciudad de Santa Clara?

Para dar respuesta a esta interrogante, nos planteamos como objetivo General, Determinar los factores de riesgo en la depresión en adultos mayores en el nivel primario de atención en la ciudad de Santa Clara, Cuba.

Como objetivos específicos proponemos,

- Identificar las variables socio-demográficas y clínicas en la población estudiada.
- Caracterizar los factores de riesgo que influyen en la aparición de la depresión en el adulto mayor.

- Establecer relación existente entre factores de riesgo y depresión en el adulto mayor.
- Validar internamente el modelo matemático obtenido para diagnosticar o clasificar, en términos probabilísticos, la depresión en adultos mayores.

2. Diseño metodológico.

Se trata de un estudio descriptivo, transversal, observacional y prospectivo, realizado con adultos mayores pertenecientes al consultorio de salud mental del policlínico "Santa Clara", del municipio de Santa Clara, en el periodo comprendido desde enero del 2018 a Julio 2021.

Para mayor claridad sobre el tipo de estudio clínico utilizado en la investigación, se describen a continuación los métodos de investigación siguientes:

Métodos teóricos: Los mismos permitieron la construcción y desarrollo de la teoría científica y el enfoque general para abordar el problema científico.

Los métodos teóricos de la Ciencia, que fueron utilizados son: el histórico – lógico, el analítico – sintético y el inductivo - deductivo, estos métodos permitieron el camino para alcanzar los resultados y fueron aplicados los siguientes:

- *Histórico-lógico:* Dado porque se parte de una revisión exhaustiva de toda la evolución que ha tenido el comportamiento clínico- socio demográfico de las patologías psiquiátricas con relación al desarrollo de la humanidad y sus descubrimientos.
- *Analítico-sintético:* Este método está a lo largo de la investigación, permitiendo diagnosticar y sintetizar el objeto de estudio, utilizándose desde la revisión bibliográfica, documental, hasta la formación de los aspectos teóricos fundamentales sobre el tema abordado.
- *Inductivo-deductivo:* Al generalizar los resultados de los estudios bibliográficos y documentales, que se efectuó en el desarrollo de la investigación, con lo cual se fue conformando los aspectos fundamentales de cuerpo de la Tesis, que se materializa en el orden de tratar oportunamente estas patologías, así como el actuar oportuno después de las intervenciones quirúrgicas evitando que evolucionen a formas severas y de esta manera mejorar la calidad de vida de estas personas.

Del nivel empírico:

- Observación científica, se elaborará una guía para la recolección de los datos.
- Revisión de documentos, se revisará las historias clínicas de los pacientes.
- Escala de Depresión Geriátrica de Yesayage
- Entrevista dirigida en busca de las variables

2.1 Población

La población de estudio está constituida por el total de pacientes, adultos mayores, que son atendidos en el consultorio de salud mental del policlínico "Santa Clara" del municipio de Santa Clara, en el periodo comprendido desde enero del 2018 a julio 2021.

2.2 Muestra del estudio

Dadas las dificultades impuestas por la pandemia que la población cubana y el mundo está padeciendo, se incorporaron al estudio los expedientes de pacientes con información completa y que aceptaron contestar la encuesta para evaluar, según Escala de Depresión Geriátrica de Yesavage (1986), su estado de depresión. En total se entrevistaron 324 adultos mayores, de los cuales 185 no presentaban depresión y 139 sí presentaban la patología, que clasificaba entre moderada o severa, que pretendemos estudiar y de la cual buscamos los factores de riesgo asociados. Por lo anterior se trata de una muestra no aleatoria, para la cual se establecieron como criterios de inclusión que el paciente sea adulto mayor, que sea atendido en el policlínico y que tenga información completa y fiable en su expediente. Las características numéricas de la población, se estimaron utilizando una técnica de remuestreo denominada Bootstrap (Efron, Bradley, 1979), a partir de la cual es posible estimar el error de estimación e intervalos de confianza.

Como criterios de exclusión se estableció lo siguiente: Pacientes que presentaban diagnóstico de Depresión mayor conocida previamente, según el Manual Diagnóstico y Estadístico de Trastornos Mentales, quinta edición(DSM V), enfermedad neurológica (Accidente cerebrovascular Enfermedad de Alzheimer, Demencia frontotemporal, Enfermedad de Parkinson, y otras enfermedades neurodegenerativas), con diagnóstico de Esquizofrenia, abuso o dependencia de sustancias, y/o alcohol, según DSM V, pacientes bajo tratamiento farmacológico antidepresivo o duelo en curso.

2.3 Instrumento de recogida de información.

La información se obtuvo por el método de encuestas a través de entrevista y la aplicación de cuestionario elaborado al efecto. Este procedimiento se realizará en horario de la mañana, en el hogar del participante, de forma individual y bajo la supervisión del investigador.

La Escala de Depresión Geriátrica (G.D.S), es un instrumento de amplio uso en la investigación gerontológica, y cuenta con los atributos de fiabilidad y validez requeridos para su uso, sin embargo, con la finalidad de verificar la fiabilidad, calcularemos su consistencia interna. Fue diseñada por especialistas del Centro de Investigaciones sobre Longevidad, Envejecimiento y Salud (CITED) para ser aplicada a los ancianos. Se tomaron una serie de factores de riesgo biológicos y psicosociales de discapacidad física de la literatura. El objetivo de su aplicación es la comprobación de la presencia o no de depresión en adultos mayores; debe aplicarse por un profesional de la salud, y el tiempo requerido para su aplicación es breve (aproximadamente 5 min).

Se realizará una única entrevista individual sin tiempo delimitado, que oscilarán entre 30 minutos y una hora de duración. Durante esta entrevista se aplicarán dos técnicas: Una entrevista semiestructurada y la escala de depresión geriátrica de Yesagave acortada, se trata de una escala específica para los adultos mayores, consta de 15 preguntas de fácil respuesta sí o no, la cual evalúa la presencia de depresión en el anciano, así como el grado en el que se encuentra ésta en caso de padecerla.

2.4 Definición de variables.

El estudio incluyó como variable dependiente la presencia de Depresión. El diagnóstico de esta entidad se obtuvo al aplicar el instrumento G.D.S a los entrevistados.

Como se sabe, el cuestionario original incluye 15 ítems, a los cuales el adulto mayor debe contestar sí o no, asignando el código 1 a la respuesta sí y 0 a no, para su tratamiento computacional. Por lo anterior, si un adulto mayor contesta "Si" a todos los ítems del cuestionario, se obtendría la puntuación máxima de 15 puntos, pero, por el contrario, si contestara "No" a todos los ítems, obtendría una puntuación de cero. La puntuación 15 equivale al máximo nivel de depresión y el cero equivale a que no presenta ningún síntoma de depresión. Debido a que pretendemos, con esta escala, clasificar a los adultos mayores como deprimidos o no, estableceremos como punto de corte en la escala 4 puntos. Nuestra clasificación quedaría de la siguiente manera:

Si el resultado es de 0 - 4 puntos se considera como Normal (No hay Depresión); si el resultado es mayor de 4 puntos, entonces clasifica como patológico (Hay Depresión).

Con esta clasificación obtenemos una variable dicotómica con las categorías Depresión (D) y No deprimido (ND), con lo que tenemos una variable del tipo Bernoulli, que puede ser expresada en términos de "éxito" y "fracaso", con probabilidad de éxito p; es decir, p es la probabilidad de que, al seleccionar un adulto mayor, clasifique como deprimido. Ahora, si codificamos la variable como D = 1, cuando ocurre "éxito" y como D=0 cuando ocurre "fracaso", y repetimos el experimento en cada uno de los adultos mayores a los cuales se les aplicó el cuestionario, entonces se configura una variable del tipo Binomial, si las observaciones fueron independientes.

Las variables que consideramos pueden estar asociadas a la depresión se indican en seguida y usualmente se denominan como variables predictoras o covariables, aunque también suelen llamarse variables independientes, nombre que en este contexto no es del todo apropiado, puesto que deseamos establecer si existe asociación entre la depresión y los factores de riesgo o covariables, por lo que serán independientes solo si no existe asociación.

Las covariables consideradas en el presente estudio serán Sexo (M= Mujer, H= Hombre), Edad (en años), Presencia de enfermedades crónicas (Si= 1, No= 0), Ausencia de confidente (Si= 1, No= 0), Inadaptación a la jubilación (Si= 1, No= 0), Condiciones económicas del entorno familiar (Buena, regular, Mala), Consumo de alcohol u otras sustancias (Si= 1, No= 0), APF de trastornos mentales (Si= 1, No= 0), Antecedentes y eventos negativos en la vida (Si= 1, No= 0), Estilos de afrontamiento (Evitativo, Activo), Personalidad premórbida con

rasgos patológicos (Si= 1, No= 0), Red de apoyo (Si= 1, No= 0), Reinserción laboral (Buena, regular, Mala), Vive solo (Si= 1, No= 0), Viudez (Si= 1, No= 0).

2.5 Modelo de regresión logística.

El modelo puede escribirse de manera compacta, dado que todas las variables explicativas son categóricas, excepto la edad, y binarias consideradas como factores de riesgo.

$$P[D = 1|x_{i1}, \cdots, x_{i16}] = \frac{e^{L_i}}{1 + e^{L_i}} = \frac{1}{1 + e^{-L_i}}$$

donde

$L_i = \beta_0 + \sum_{j=1}^{16} \beta_j \, x_{ij}, j = 1, \cdots, 16; \;\; i = 1, \cdots, n.$

$\beta_0 = constante$ y β_j es el j-esimo parámetro asociado a la i-ésima variable explicativa x.
L_i es un modelo de regresión, que en el modelo de regresión logística recibe el nombre de componente sistemática o regresor lineal, cuyos parámetros $\beta = (\beta_0, \beta_1, \ldots, \beta_{16})$ deben ser estimados.

Al ejecutar el procedimiento en software computacional, deberá realizarse un contraste de bondad de ajuste del modelo, lo cual se consigue con una prueba de razón de verosimilitud y de Hosmer y Lemeshow, por lo cual debe observarse si el valor de *-2log de la verosimilitud* es "grande" (para el caso de un valor de α=0.05, $\chi^2 = 3.83$ para 1 grado de libertad), puesto que es evidencia en contra de la hipótesis nula, lo mismo ocurre si el p-valor < 0.05 = α, para la prueba de Hosmer y Lemeshow en el último paso del método, con lo cual se comprueba el ajuste global del modelo.

Se aplicó la prueba de Homogeneidad basada en la distribución Chi cuadrado para identificar asociaciones entre variables cualitativas o categorizadas. Como resultado de la misma se mostró el valor de su estadígrafo (χ^2), así como la significación asociada al mismo (p).

De acuerdo al valor de p se clasificó la relación o asociación en *significativa* si p es menor que 0.05, *no significativa* si p es mayor que 0.05.

Dado el número de variables estudiadas se inició el tratamiento de los datos con el análisis de tablas de contingencia (prueba de homogeneidad basada en la Distribución de Chi cuadrado o χ^2), con un nivel de significación del 5%. Las variables que no resultaron eliminadas se incluyeron en un segundo análisis, que consistió en la aplicación del un Modelo de Regresión Logística con respuesta dicotómica, por el método paso a paso.

La aplicación del Modelo de Regresión Logística se realizó, previa comprobación de la ausencia de colinealidad entre variables independientes. Por cada variable independiente se contó con un con un mínimo de diez individuos por cada evento de la variable dependiente.

El ajuste del Modelo a los datos se verificó a través del estadígrafo de Hosmer y Lemeshow. La significación estadística asociada al estadígrafo de la prueba fue mayor de 0.05, por tanto, se consideró que el modelo se ajusta a los datos. Además, se estimaron los OR puntuales y por intervalos de confianza para cada variable independiente.

Se utilizo la Técnica de Validación Interna conocida como *Validación Aparente*: esta consiste en evaluar el rendimiento del modelo usando los mismos datos que han sido utilizados en el desarrollo del mismo. Este método *usa todos los datos disponibles* para la creación y validación del modelo.

Con las probabilidades obtenidas se construyó una Curva ROC (Receiver Operating Characteristic Curve). La evaluación de la capacidad predictiva del modelo se realizó por el examen visual de la curva. El AUC fue de 0.854 (cuando los valores se encuentran entre 0.75 y 0.9: la capacidad predictiva es buena).

3. Resultados y discusión.

Para cumplir los objetivos planteados en la investigación, se obtuvo la información de las diferentes fuentes empleadas, la cual fue procesada en el software computacional SPSS y cuyos resultados presentamos.

*Tabla 1: Comparación de variables: Sexo * Depresión*

		Depresión		
		No	Si	Total
Sexo	Mujer	98	89	187
		52.4%	47.6%	100.0%
	Hombre	87	50	137
		63.5%	36.5%	100.0%
Total		185	139	324
		57.1%	42.9%	100.0%

Según observamos en la tabla 1, el total de casos de personas con depresión, son. Efectivamente, 139 casos, que representan el 42.9% de la muestra estudiada, lo que significa que, por cada deprimido, en sus distintos grados, hay 1.33, en promedio, que no lo está.

Se realizó un análisis de tablas de contingencia con la finalidad de identificar a las variables que son estadísticamente significativas para explicar la depresión , de las cuales se calculó el OR (Razón de Odds, Razón de momios o Razón de ventajas) y encontramos que las variables que pueden ser candidatas a incorporarse al modelo de regresión logística son las resumidas en la tabla 2, aunque, como describiremos en la ejecución del procedimiento paso a paso de Wald, debido a que se trata de un algoritmo, la selección se hace mediante mecanismos probabilísticos para la inclusión o exclusión de una variable del modelo.

Para los resultados de la tabla 2, la prueba estadística que se ha utilizado para determinar la significación estadística de una variable que será candidata a incluirse en el modelo, es la prueba de independencia utilizando un estadígrafo χ^2, fijando un nivel de significación $\alpha = 0.05$ para rechazar la hipótesis de independencia de las variables cuando $\chi_0^2 \geq \chi_{1,\alpha}^2$ o cuando p-valor < 0.05.

Tabla 2: Significación estadística de las variables comparadas contra Depresión (No/Si) y OR.

2.6 Variables	Sig. asintótica (bilateral)	OR	Intervalo de confianza al 95%		1/OR	Interpretación
			Inferior	Superior		
Sexo(M/H)	0.046	0.63	0.40	0.99	1.58	La ventaja a favor de depresión, es hasta 58% superior si el paciente es mujer
Presencia de enfermedades crónicas (Si/No)	0.001	2.57	1.47	4.49		La ventaja a favor de D, es hasta 157% superior para los adultos mayores que tienen enfermedades crónico degenerativas.
Pérdidas (Si/No)	0.004	0.49	0.30	0.80	2.04	La ventaja a favor de depresión, es hasta el 100% superior si el paciente no ha sufrido pérdidas.
Ausencia de confidente (Si/No)	0.03	1.70	1.05	2.75		La ventaja a favor de D, es hasta del 70% superior para los adultos mayores sin confidente.
Antecedentes y eventos negativos en la vida (Si/No)	0.00	2.77	1.69	4.54		La ventaja a favor de D, es hasta 177% superior para los que tienen Antecedentes y eventos negativos en la vida.
Estilos de afrontamiento (Evitativo / Activo)	0.00	0.066	0.034	0.130	15.15	La ventaja a favor de D, es hasta 1400% superior para los pacientes con estilo de afrontamiento Activo.
Personalidad premórbida con rasgos patológicos (Si/No)	0.00	0.31	0.18	0.53	3.22	La ventaja a favor de D, es hasta 222% superior si el paciente no tiene Personalidad premórbida con rasgos patológicos.
Red de apoyo (Si/No)	0.012	0.41	0.20	0.84	2.42	La ventaja a favor de D, es hasta 142% superior si el paciente no cuenta con una red de apoyo.
Reinserción laboral (Si/No)	0.02	0.54	0.32	0.89	1.87	La ventaja a favor de depresión es hasta 87% superior para los que no tuvieron oportunidad de reinserción laboral
Vive solo (Si/No)	0.00	3.02	1.82	5.01		La ventaja a favor de D, es hasta 200% superior para los pacientes que viven solos.

Fuente: Elaboración propia a partir de la salida de SPSS.

*En el caso de un OR entre 0 y 1, halló el recíproco, lo que equivale a permutar las filas R1 con R2, con la finalidad de interpretar en términos de valores mayores que 1 y depresión Si. Recordar que un OR cercano a 1, es evidencia de independencia entre variables y no es de utilidad para nuestro propósito, razón por la cual se omiten de la tabla (buscamos asociación).

Cuando el OR es mayor que 1 significa que la variable comparada con depresión es un Factor de Riesgo, como en el caso de las variables presentadas en la tabla resumen. Observemos que Sexo es clasificada como una variable que contribuye a explicar la depresión, aunque habrá que tratarla con cautela, toda vez que su p-valor

es muy cercano al valor de prueba (0.05), aunque ligeramente menor (0.046), la diferencia es de apenas cuatro milésimas (0.004). Habrá que decidir si se incorpora al modelo o no, a partir de este hecho, dado que es posible que el sexo no tenga influencia en la depresión, es decir, ocurre por igual en ambos grupos.

Ahora, el resultado del procedimiento para regresión logística binaria se presenta y discute a continuación, en el cual se comprueba la idoneidad del modelo de regresión logística binaria mediante la prueba de ómnibus sobre los coeficientes del modelo, cuyo p_valor es cero, lo que indica que los coeficientes son significativamente distintos de cero y que los datos se pueden ajustar al modelo. Lo anterior se confirma al observar la prueba de razón de verosimilitud y de Prueba de Hosmer y Lemeshow, en el último paso del método de la salida computacional.

La tabla de clasificaciones correctas también puede tomarse como un elemento de información para la idoneidad del modelo, en nuestro caso, el modelo clasifica correctamente, aproximadamente, el 78% de los casos

Tabla 3: Tabla de clasificación[a]

	Observado		Pronosticado		
			Depresión		Porcentaje correcto
			Si	No	
Paso 6	Depresión	Si	96	43	69.1
		No	29	156	84.3
	Porcentaje global				77.8
a. El valor de corte es 0.500					

Tabla 4: Variables en la ecuación

		B	E.T.	Wald	gl	Sig.	Exp(B)	I.C. 95% para EXP(B)	
								Inferior	Superior
Paso 6[a]	perdidas(1)	1.108	.324	11.706	1	.001	3.027	1.605	5.708
	inadaptacionjubilacion(1)	-.655	.332	3.892	1	.049	.519	.271	.996
	eventosnegativos(1)	-1.634	.351	21.694	1	.000	.195	.098	.388
	afrontamiento(1)	3.178	.404	61.765	1	.000	24.006	10.866	53.035
	reddeapoyo(1)	.936	.467	4.009	1	.045	2.550	1.020	6.373
	vivirsolo(1)	-1.264	.331	14.604	1	.000	.283	.148	.540
	Constante	-2.080	.587	12.541	1	.000	.125		
a. Variable(s) introducida(s) en el paso 6: inadaptacionjubilacion.									

La tabla 4 muestra las variables que deberán incluirse en el modelo de regresión lineal, aunque debemos observar que la constante no estadísticamente significativa para ser considerada en el modelo, aunque si la omitimos, la tasa de clasificaciones correctas no se ve modificada.

Por lo anterior, el modelo de regresión logística binaria deberá formularse de la siguiente manera:

$$P[Y = 0|L_i] = \frac{1}{1 + e^{-L_i}}$$

donde,

$$L_i = -2.08 + 1.108Perdidas - 0.655InadaptacionJubilacion - 1.634Eventosnegativos + 3.178Afrontamiento + 0.936Redeapoyo - 1.264Vivirsolo$$

Entonces, con el modelo encontrado es posible clasificar a nuevos pacientes con solo medir las variables que se incorporan en el presente regresor lineal, que, en nuestro caso, son variables del tipo dicotómico, es decir, presencia o ausencia de la condición.

Supongamos que un paciente ha sufrido pérdidas, entonces el valor de pérdida es 1, se ha adaptado a la jubilación (0), ha experimentado eventos negativos (1), tiene estilo de afrontamiento evitativo (0), tiene buena red de apoyo (1) y vive solo (1). Sustituimos en el modelo estos valores y se obtiene P=0.05, lo que significa que el paciente clasifica en el grupo de abajo, es decir, en el grupo cuyas probabilidades son cercanas a cero, es decir, clasifica en el grupo de los no deprimidos, por el contrario, si su probabilidad es cercana a 1, significa que clasifica en el grupo de los deprimidos. En una hoja de cálculo puede programarse el modelo obtenido y hacer los cálculos manualmente con solo introducir los valores de las variables predictoras resultantes del modelo, que en nuestro caso, son variables categóricas con las categorías "*presencia*" o "*ausencia*" de la condición, es decir, 1 o 0.

$$L_i = -2.08 + 1.108(1) - 0.655(0) - 1.634(1) + 3.178(0) + 0.936(1) - 1.264(1) = -2.934$$

Entonces,

$$P[Y = 1|L_i = -0.439] = \frac{1}{1 + e^{-L_i}} = 0.05$$

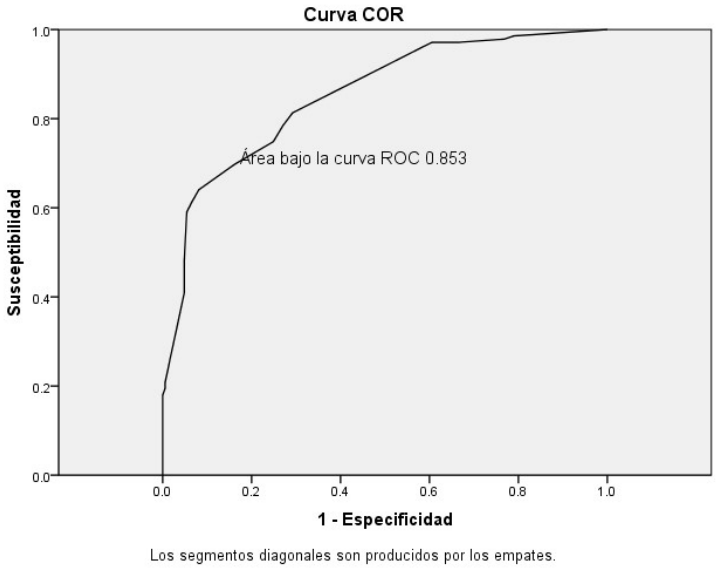
Ahora, según lo establecido en la teoría, la exponencial de los parámetros $Exp(\beta) = OR$, por lo que podemos interpretar de forma similar a como lo hemos hecho para el caso de las tablas de contingencia, aunque las variables candidatas que se han descartado en el modelo son *sexo, presencia de enfermedades crónicas* y *personalidad premórbida con rasgos patológicos*, aunque se incorporó *Inadaptación a la jubilación*.

Tabla 5: Interpretación de las OR del modelo

Variable	$Exp(\beta) = OR$	1/OR	Interpretación
perdidas (1)	3.027		La ventaja a favor de D es superior en 200% para aquellos pacientes que han sufrido pérdidas.
inadaptacionjubilacion(1)	.519	1.93	La ventaja a favor de D es superior hasta en un 93% para los pacientes que no se adaptaron a la jubilación.
eventosnegativos(1)	.195	5.13	La ventaja a favor de D, es hasta 4 veces superior para aquellos pacientes que han enfrentado eventos negativos.
afrontamiento (1)	24.006		La ventaja a favor de D es superior es hasta 23 veces para aquellos pacientes con estilo de afrontamiento activo.
reddeapoyo(1)	2.550		La ventaja a favor de D es hasta 150% superior para los pacientes que tienen mala red de apoyo.
vivirsolo(1)	.283	3.53	La ventaja a favor de D es superior 253% para aquellos pacientes que viven solos.

Constante			

Otro recurso para comprobar la idoneidad del modelo es la curva ROC, como se dijo antes, por lo que se presenta a continuación y en la que se confirma que el modelo tiene buena capacidad predictiva (cuando los valores se encuentran entre 0,75 y 0,9: la capacidad predictiva es buena).



4. Conclusiones.

En el estudio se constató un mayor número de pacientes con riesgo de padecer depresión cuando presentan eventos negativos en su vida y viven solos. Las variables sociodemográficas y clínicas tomadas en cuenta para este estudio fueron edad, sexo y enfermedades crónicas, las cuales, según nuestros resultados, no tienen un impacto significativo en la depresión. Uno de los factores de riesgo relacionados con la depresión en el adulto mayor fue la presencia de eventos negativos caracterizados por el fallecimiento de algún familiar de primera línea, así como vivir solo. El modelo obtenido permite, a partir de los perfiles de riesgo individuales, el diagnóstico de depresión en la población de ancianos estudiada. Entonces, si observamos las variables *pérdidas, inadaptación a la jubilación, eventos negativos en su vida, estilo de afrontamiento de pérdidas, red de apoyo familiar y vivir solo,* con las categorías presencia (1) o ausencia (0) de la condición, es posible hallar la probabilidad de que un paciente clasifique como deprimido o no.

Bibliografía

Buiques C, P. S.-R. (2015). The Relationship between Depression and Frailty síndrome: a systematic review. *Aging Ment Health.*, 762-772.

Calderón M., D. (2018). Epidemiology of depression in the elderly. *Revista Médica Herediana*, 182-191.

Díaz-Villa, B., & González-González, C. (2012). Actualidades en neurobiología de la depresión. *Revista Latinoamericana de Psiquiatría*, 106-115.

Freitas C, D. S. (2016). Risk of diabetes in older adults with co-occurring depressive symptoms and cardiometabolic abnormalities: Prospective analysis from the english longitudinal study of ageing. *PLoS ONE*, 1-16.

González-Menéndez, R. (2008). *Paciente psicogeriátrico y su atención específica.* La Habana: Editorial Ciencias Médicas.

Luna-Reyes, T., & Vilchez-Hernández, E. (2017). Depresión: Situación actual. *Revista de la Facultad de Medicina Humana*, 78-85.

OMS. (2002). *II Asamblea Mundial de las Naciones Unidas sobre el Envejecimiento: Envejecimiento activo.* Madrid: OMS.

OPS. (2013). *Depresión, Guía de Diagnóstico y manejo.* Washington, DC: OPS.

Tello-Rodríguez, T., & Varela-Pinedo, L. (2016). Fragilidad en el adulto mayor: detección, intervención en la comunicad y toma de desiciones en el manejo de enfermedades crónicas. *Rev Perú Med Exp Salud Publica.*, 33-34.

Vaughan, L., Corbin, A., & Goveas, J. (2015). Depression and frailty in later life: A systematic review. *Clin Interv Aging.*, 1947-1958.

Wolff L, C. A. (2010). Prevalencia, factores de riesgo y manejo de la depresión en pacientes con infección por VIH: Revisión de la literatura. *Revista chilena de infectología.*, 65-72.

Capítulo 9 pp 135-143

GENERACIÓN DE DATOS SINTÉTICOS USANDO REDES BAYESIANAS CONSERVANDO LA MATRIZ DE CORRELACIÓN

Salgado Guzmán, Oscar Rene· Sandoval Solís, María de Lourdes, Rivera Martínez, Marcela and Marcial Castillo, Luis René
Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla, Puebla, México

Resumen

Cuando se desea conservar la confidencialidad de datos, es importante generar datos sintéticos, manteniendo las propiedades de las correlaciones entre los atributos, así como su distribución de probabilidad de cada uno de ellos. En el presente trabajo, se usa el software DataSynthesizer [4] para generar datos sintéticos usando redes bayesianas conservando la matriz de correlación para la base de datos Adult income[8]. Se muestra que el software conserva la matriz de correlación y la distribución de los atributos para diferentes parámetros, y que al aumentar el número de datos sintéticos generados se acerca más a las propiedades de los datos originales.

Keywords: Datos Sintéticos, DataSynthesizer, Ruido de Laplace, Redes Bayesianas, Inteligencia Artificial.

1 Introducción

Para entrenar módulos de inteligencia artificial se utilizan base de datos, pero a veces la base de datos tiene datos de mala calidad, también cuando se requiere trabajar con bases las cuales cuentan con pocos datos o donde las bases de datos tienen información personal y privada de las personas que podrían comprometer la confidencialidad del individuo, como en el área de la salud o en el área financiera, es entonces donde los datos sintéticos se generan presentando una solución que se puede trabajar con base de datos sin comprometer información confidencial.[1][2]

Los datos sintéticos son información que se genera de forma artificial para sustituir los datos históricos reales con el fin de entrenar modelos de inteligencia artificial. Se recurre a este tipo de datos porque los datos reales son insuficientes (o de mala calidad) e imposibilitan el uso de técnicas de inteligencia artificial. Los datos sintéticos permiten conservar la privacidad. [3]

Se requiere trabajar con datos sintéticos para ello es necesario generarlos, en este documento se experimenta con redes bayesianas las cuales usando la matriz de correlación se puede identificar el grado de dependencia entre las variables, las redes bayesianas permiten predecir datos, ya que una red bayesiana es mejor que una red neuronal para la inferencia estocástica. Consecuentemente, se explica el uso de DataSynthesizer para generar datos sintéticos y crear conjuntos de datos que se parecen a la muestra, esto se aplica a la base de datos original Adult data set.

A continuación, se presenta un resumen sobre la teoría de las redes bayesianas, para después introducir el software de Datasynthesizer. Más adelante se expone la experimentación tanto en la base de datos original, así como en los datos sintéticos, también se muestra la descripción del trabajo. Para posteriormente presentar resultados y finalizar con las conclusiones después de hacer un análisis de los resultados. En el final del documento se dan las referencias de los documentos y artículos consultados.

2 Estado del arte

Existe un gran número de trabajos que aportan al tema de los datos sintéticos, así como un considerable número de herramientas para su generación. En este proyecto se consultaron artículos que se centran en la generación de datos sintéticos y que, además, proponen herramientas y software para obtener dichos datos.

En [4] Fida y Mahmoud exponen los métodos para obtener datos sintéticos efectivos. En su trabajo se ocupan de evaluar el efecto de varias configuraciones de generación y uso de datos sintéticos sobre la utilidad de los datos sintéticos generados y sus modelos derivados; el efecto del preprocesamiento de datos en la utilidad de los datos sintéticos generados. Además de si compartir los resultados preliminares del aprendizaje automático puede mejorar los modelos de datos sintéticos. Abordan una investigación empírica de datos sintéticos generados a partir de generadores de datos sintéticos: Synthetic Data Vault, DataSynthesizer y SynthPop.

En [2] se hace un enfoque a la utilidad de los datos sintéticos para preservar la privacidad de los datos en el área de la salud. Generan datos sintéticos mediante tres herramientas que aplican enfoques en CART (Arboles de clasificación y regresión), paramétricos y en redes bayesianas. El paquete de R Synthpop, desarrollado por Nowak, ofrece una implementación disponible de las técnicas generadoras de datos sintéticos basados en paramétricos y en CART. El software DataSynthesizer, desarrollado por Ping, proporciona una implementación de generación de datos sintéticos basado en redes bayesianas. En la parte de los resultados nos dicen que el 92% de los modelos entrenados con datos sintéticos tienen menor precisión que los entrenados con datos reales, aunque la diferencia es pequeña.

En el trabajo se Kuchin y Yakunin [1] se encargan de evaluar el funcionamiento de algoritmos de machine learning (ML), para ello generaron un conjunto de datos sintéticos. Evalúan el contenido, calidad y formato de los datos, dado que estas características afectan los procesos de los algoritmos de ML. Con la utilización de redes neuronales feedforward, algoritmos k-nearest neighbor y máquinas de soporte vectorial (SVM) se dan a la tarea de clasificar los datos y medir la precisión.

Ping expone en [5] un artículo más detallado sobre cómo se generan los datos sintéticos, específicamente en el software DataSynthesizer, el cual preserva la privacidad de los datos confidenciales. Ofrecen una descripción e ilustración de su manejo en un contexto científico, puesto que los proyectos de ciencias sociales y de la salud requieren sensibilidad cuando los datos son compartidos.

A continuación, se describen:

- Synthetic Data Vault: Modela la función de distribución acumulativa F de la población a partir de la muestra. [4]
- Data Synthesizer: Se trata de una técnica de síntesis de datos basada en redes bayesianas desarrollada en Python. Captura la estructura de correlación subyacente entre los diferentes atributos mediante la construcción de una red bayesiana.[4]
- Synthpop paramétrico y no paramétrico: una técnica de síntesis de datos paramétrica y no paramétrica basada en árboles de decisión. Utiliza los algoritmos del método CART. Se implementa en R. [4]

3 Redes Bayesianas

3.1 Teoría de probabilidad

A continuación, se van a presentar algunos conceptos de probabilidad para entender la regla de bayes

Estadística Bayesiana.
Conjunto de herramientas que se utiliza en un tipo especial de inferencia estadística que se aplica en el análisis de datos experimentales en muchas situaciones prácticas de ciencia e ingeniería.

Regla de bayes.
Si los eventos Y_1, Y_2, …, Y_k representan una partición del espacio muestral S, donde P(Y_i) ≠ 0 para i = 1, 2, …, k, entonces, para cualquier evento X en S, tal que P(X) ≠ 0.

$$P(Y_r|X) = \frac{P(Y_r \cap X)}{\sum_{i=1}^{k} P(Y_i \cap X)} = \frac{P(Y_r)P(X|Y_r)}{\sum_{i=1}^{k} P(Y_i)P(X|Y_i)} \; para\; r = 1, 2, \dots, k. \qquad (1)$$

La regla de Bayes, un método estadístico llamado bayesiano, ha adquirido muchas aplicaciones. En el capítulo 18 de [10] se introduce a método bayesiano.

3.2 Teoría de grafos

Para entender lo que es una red bayesiana es necesario comprender algunos conceptos de parte de la teoría de grafos. Las siguientes definiciones nos aportaran lo básico para entender el concepto de red bayesiana.

Las redes bayesianas son una representación gráfica de dependencias para razonamiento probabilístico, en la cual los nodos (X) representan variables aleatorias y los arcos (A) representan relaciones de dependencia directa entre las variables.

Las redes bayesianas modelan un fenómeno mediante variables y las relaciones de dependencia entre ellas. Con la inferencia bayesiana se puede estimar la probabilidad posterior de las variables no conocidas, en base a las variables conocidas.

Definición:
Una red Bayesiana es una 4-tupla (G, f_x, Q, Θ), que representa una distribución de Probabilidad Conjunta donde:

- (G, f_x, Q) es una red causal.
- G es un dígrafo acíclico.
- El conjunto X de nodos de G es un conjunto $\{x_1 \mid i \leq n\}$, de variables aleatorias con r estados posibles.
- Θ es el conjunto $\{\Theta_i \mid i \leq n\}$ y $\Theta_i = \{P(x_i = k|\pi_i = j) \mid k \in Q$ y j es una configuración de los padres de $x_i\}$ donde P(x_i = k| π_i = j), denota la probabilidad de que el estado x_i sea k, dado que la configuración de padre es j.

Una red bayesiana representa en forma gráfica las dependencias e independencias entre variables aleatorias, en particular las independencias condicionales.

4 Ruido

El ruido (Noise) se agrega a la distribución de datos para conservar la privacidad, el objetivo es que la probabilidad de obtener algún resultado de los datos sea parecida a la que se podría obtener de otro conjunto de datos siendo distintos del original en un elemento. Con el software se implementa un mecanismo

diferencialmente privado, agregando ruido controlado a las distribuciones aprendidas. El ruido se agrega con la distribución de Laplace.

El mecanismo de Laplace utiliza la función de distribución de Laplace para introducir el ruido deseado a los datos para que cumpla ε-Differential Privacy. La distribución de Laplace es la distribución de la diferencia de dos variables aleatorias independientes con distribuciones exponenciales idénticas.

5 DataSynthesizer

DataSynthesizer (DS) es un sistema integral que toma un conjunto de datos privados como entrada y genera conjuntos de datos sintéticos, que simulan un conjunto de datos determinado. El sistema está implementado en Python 3[12]. Su objetivo es facilitar las colaboraciones entre científicos de datos y propietarios de datos confidenciales. Aplica técnicas de Privacidad Diferencial para lograr una fuerte garantía de privacidad.

Captura la estructura de correlación subyacente entre los diferentes atributos mediante la construcción de una red bayesiana, Las cadenas no categóricas permiten a DS generar cadenas aleatorias durante la etapa de generación de datos. Esta característica permite a DataSynthesizer crear conjuntos de datos que se parecen a la muestra real al incluir datos de cadenas sintéticas como nombres artificiales e identificaciones.

DataSynthesizer puede operar en tres modos:

- Correlation mode: Construye una red bayesiana diferencialmente privada que captura la estructura de correlaciones entre atributos y, a continuación, extrae muestras de este modelo para construir el conjunto de datos resultante.

- Independent attribute mode: En este modo se obtiene un histograma para cada atributo, se agrega el ruido al histograma para conseguir la privacidad diferencial y se extraen muestras para cada atributo. Se usa cuando el modo de atributos correlacionados es demasiado caro computacionalmente o cuando no hay datos suficientes para derivar un modelo razonable

- Random mode: Simplemente genera valores aleatorios consistentes con el tipo para cada atributo, es para casos de datos extremadamente sensibles.

5.1 Módulos

DataDescriber: Elaborar un resumen de datos.

El conjunto de datos de entrada es procesado primero por el módulo DataDescriber, guarda una descripción del conjunto de datos en un archivo JSON.

DataDescriber investiga los tipos de datos, las correlaciones y las distribuciones de los atributos en el conjunto de datos privados, y elabora un resumen de los datos, añadiendo ruido a las distribuciones para preservar la privacidad.

Primero el conjunto de datos de entrada es procesado por el módulo DataDescriber. Los dominios y las estimaciones de las distribuciones de los atributos se infieren y se guardan en un archivo de descripción del conjunto de datos.

DataDescriber requiere parámetros para tener un mejor ajuste de la descripción de datos, uno de estos parámetros es el umbral categórico, como cualquier umbral, puede ser difícil de establecerlo de forma que refleje las preferencias del usuario.

Otro parámetro importante es épsilon, un parámetro de privacidad diferencial. Significa que eliminar una fila del conjunto de datos de entrada no cambiará la probabilidad de obtener el mismo resultado más que una diferencia multiplicativa de exp(épsilon). Se aumenta el valor de épsilon para reducir los ruidos inyectados. Establecer épsilon=0 para desactivar la privacidad diferencial.

DataGenerator: Generar un conjunto de datos sintéticos a partir del resumen.

DataGeneretor toma muestras de la distribución de frecuencias de los valores calculados con DataDescriber para obtener los datos sintéticos.

En este módulo se selecciona el modo de operar del software, ya sea Independent attribute mode (modo de atributo independiente), Correlation mode (modo correlacionado) o random mode (modo aleatorio); Cuando se invoca el modo aleatorio DataGenerator genera valores aleatorios de tipo coherente para cada atributo. En el modo de atributo independiente extrae muestras de gráficos de barras o histogramas mediante muestreo uniforme. Y cuando se hace el llamado al modo correlacionado se muestrean los valores de los atributos conservando la matriz de correlaciones de los datos generados a partir de la red bayesiana.

ModelInspector: Inspección y comparación de conjuntos y resúmenes de datos.

ModelInspector muestra una descripción intuitiva del archivo description.json de datos calculado por DataDescriber, lo que permite al propietario de los datos evaluar la precisión del proceso de generación y ajustar los parámetros, si lo desea.

Proporciona varias funciones integradas para inspeccionar la similitud entre el conjunto de datos privados de entrada y el conjunto de datos sintéticos de salida. El propietario de los datos puede comprobar rápidamente si las tuplas del conjunto de datos sintéticos son detectables inspeccionando y comparando las 5 primeras y las 5 últimas tuplas de ambos conjuntos de datos.

Algoritmos.

La distribución condicional es construida de acuerdo con Algorithm 1 de [6]. En este algoritmo se encarga de agregar el ruido Laplace a los atributos y el Algorithm 2 de [5] describe el proceso de la generación de datos.

Algorithm 1: NoisyConditionals (D, N, k): regresa P *.

1.Inicializar variable para guardar la distribución condicional de salida, P* = 0;
2.for i = k +1 to d do: #i = k+1 hasta número de atributos.
3. materializar la distribución conjunta Pr[X_i, $\prod_i$]; #Se genera el espacio para la distribución condicional del hijo X_i.
4. se genera la privacidad diferencial Pr*[X_i, $\prod_i$] agregando el ruido de Laplace;
5. establecer los valores negativos de la privacidad diferencial Pr*[X_i, $\prod_i$] a 0 y normalizar;
6. obtener la distribución condicional con ruido Pr*[X_i | $\prod_i$] de la distribución en Pr*[X_i, $\prod_i$]; agregarlo a P*;
7.for i = 1 to k do:
8. obtener la distribución condicional con ruido Pr*[X_i | $\prod_i$] de Pr*[X_{k+1}, $\prod_{k+1}$]; se agrega a P*;
9.return P *;

Algorithm 2: DataGenerator (n, M, S, A_u. , *s)*

Se requiere: número de tuplas n a generar, modo M, descripción del conjunto de datos S, atributos uniformes A_u, semilla s.

1 Establecer semilla = s para el generador de números pseudoaleatorios.
2 if *M* es modo de atributos independientes (independent attribute mode) then:
3 Leer todos los atributos A de *S*.
4 for X ∈ A do #Se hace recorrido por los atributos
5 if X ∈ A_u then
6 Leer el dominio de X a partir de *S*.
7 Muestrea *n* valores uniformemente de su dominio.
8 else
9 Lee la distribución de X a partir de *S*.
10 Muestrea *n* valores de su distribución.
11 end if
12 end for
13 else if *M* es modo de atributos correlacionados (correlated attribute mode) then:
14 Leer la red bayesiana *N* a partir de *S*.
15 Muestra atributo raíz de una distribución No condicional.
16 Muestra atributos restantes de distribucion condicional.
17 end if
18 regresa el Conjunto de datos muestreados;

6 Experimentación

6.1 Recuperación de datos

Se recupera la base de datos Adult Data Set de UCI Machine Learning Repository [http://archive.ics.uci.edu/ml] [8][9]. Esta base de datos contiene información de los ingresos de adultos basándose en los datos del censo. También conocido como conjunto de datos "Census Income".

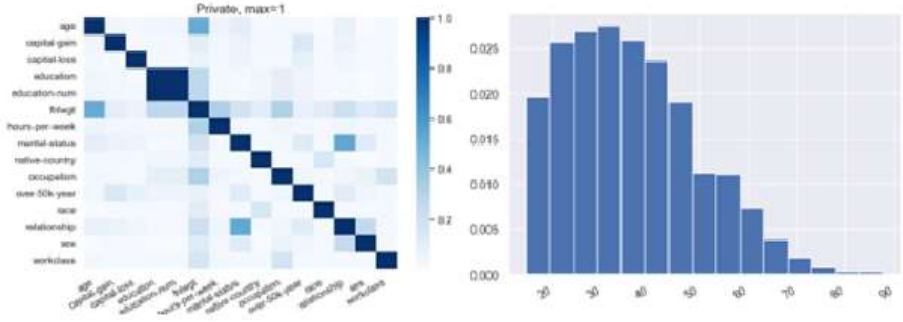


Fig. 1. Matriz de correlación de atributos de datos originales: Adult income y Distribución de probabilidad de atributo Age: Adult income

6.2 Resultados

Se generan los datos sintéticos con la biblioteca DataSynthesizer de Python, se puede encontrar el código fuente y aplicaciones en ejemplos en [7][10][11]. A continuación, se presentan algunas observaciones.

Primero se obtienen los datos sin ruido para esto se debe igualar la variable épsilon con 0. En la matriz de correlación para los datos sintéticos, no se observan grandes diferencias en los datos sintéticos generados, a pesar de que se aumenta el número de datos sintéticos. En cambio, el histograma del atributo *age*, al aumentar el número de datos sintéticos a generar se parece más su distribución a la original. También, se observa que

cuando se cambia el grado de la red bayesiana, no hay cambios sustanciales ni en la matriz de correlación ni en el histograma.

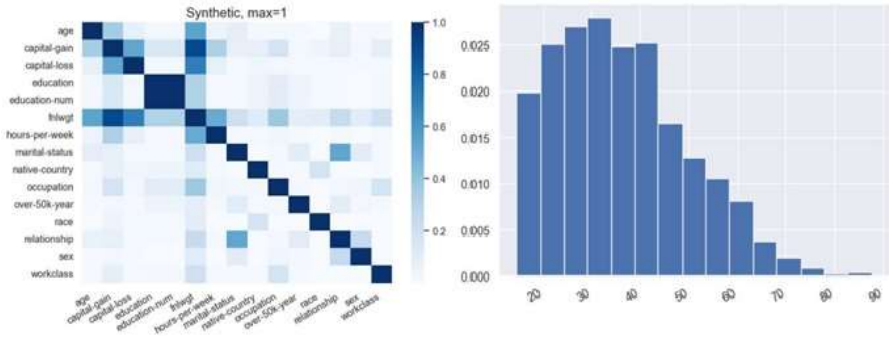


Fig. 2. Matriz de correlación de datos sintéticos y distribución de atributo Age de datos sintéticos generados sin inyectarles ruido.

Ahora, con los resultados a los cuales le agregamos valor a épsilon, es decir generar datos sintéticos con ruido.

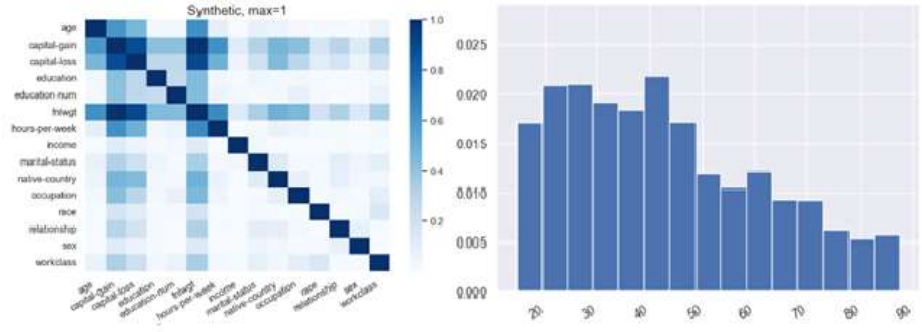


Fig. 3. Matriz de correlación de datos sintéticos y distribución de atributo Age de datos sintéticos generados con ruido (Prueba a).

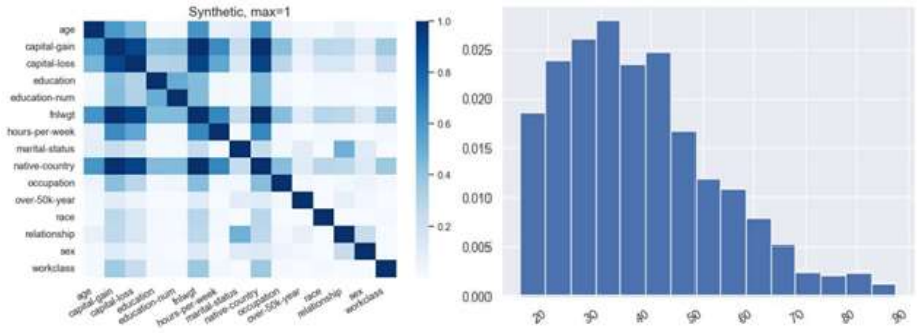


Fig. 4. Matriz de correlación de datos sintéticos y distribución de atributo Age de datos sintéticos generados con ruido (Prueba b).

6.3 Observaciones.

Por parámetro:

Número de datos sintéticos: Después de obtener resultados con diferente cantidad de datos generados, se concluye que se tiene un mejor comportamiento de los datos sintéticos entre más alto el número de tuplas a obtener.

Número de nodos: Para el número máximo de padres para cada nodo, se observa que un mejor ajuste es cuando cada nodo tiene máximo dos padres.

Umbral: De acuerdo con la experimentación se observa que un mejor comportamiento en las variables es cuando el umbral es cercano al dominio del atributo categórico con más valores diferentes.

Épsilon: Con antelación se menciona la utilidad del valor épsilon, y para el software DataSynthesizer se debe tener en cuenta que tanto ruido se desea ingresar a los datos, ya que con un ruido mayor la distribución entre datos originales y sintéticos tendrá una diferencia significativa, cuando se generan datos sintéticos con ruido para la matriz de correlación. En cambio, en el histograma de distribución se observa que hay una diferencia significativa al cambiar el valor de épsilon.

7 Conclusiones

Los datos sintéticos nos ayudan a entrenar modelos de inteligencia artificial cuando los datos son insuficientes o para no comprometer información confidencial, resultando en datos con utilidad para diferentes proyectos. Entre una variedad de software para generar datos sintéticos se tiene DataSynthesizer, un software generador de datos sintéticos, utilizando redes bayesianas para calcular las probabilidades de relación, además de conservar la matriz de correlación entre los datos.
Para este trabajo se ocupa la librería DataSynthesizer por el hecho de que es un software sencillo de instalar, a diferencia de otras opciones, esto resulta en una obtención de datos más ágil. Para conservar la matriz de correlación este software emplea redes bayesianas, por ello se hizo un repaso sobre redes bayesianas tanto en teoría de probabilidad como en teoría de grafos.

Se ejemplificó su uso con la base de datos Adult data set.

Referencias

1. Yan I. Kuchin, Ravil I. Mukhamediev & Kirill O. Yakunin | Duc Pham (Reviewing editor) (2020) One method of generating synthetic data to assess the upper limit of machine learning algorithms performance, Cogent Engineering, 7:1, DOI: 10.1080/23311916.2020.1718821
2. Debbie Rankin, Maurice Mulvenna, Michaela Black, Raymond Bond, Jonathan Wallace, Gorka Epelde (2020). Reliability of Supervised Machine Learning Using Synthetic Data in Health Care: Model to Preserve Privacy for Data Sharing. JMIR Med Inform; 8(7) :e18910. DOI: 10.2196/18910.
3. Fernando Fuentes (2022). Datos sintéticos, un recurso vital para la Inteligencia Artificial. https://www.arsys.es/blog/datos-sinteticos, Accesado el 27/09/2022
4. Dankar, F.K.; Ibrahim, M. Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. Appl. Sci. 2021, 11, 2158. DOI: 10.3390/app11052158.
5. Haoyue Ping, Julia Stoyanovich, and Bill Howe (2017). DataSynthesizer: Privacy-Preserving Synthetic Datasets. In Proceedings of SSDBM ’17, Chicago, IL, USA, June 27-29, 2017, 5 pages. DOI: 10.1145/3085504.3091117.
6. Jun Zhang and others. 2014. PrivBayes: private data release via Bayesian networks. In SIGMOD.

7. Ping, H., & Yang, K. (2020, 11 junio). DataSynthesizer. GitHub. Accesado 14/Enero/2023. de https://github.com/DataResponsibly/DataSynthesizer
8. Kohavi, Ronny. Becker Barry(1996). Adult Data Set. UCI Machine Learning Repository. https://archive.ics.uci.edu/ml/datasets/adult. Accesado el 23/11/2022.
9. Dua, D. and Graff, C. (2019). UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science. Accesado el 23/11/2022.
10. Haoyue, Ping. DataSynthesizer Usage (correlated attribute mode). correlated_attribute_mode.ipynb. https://github.com/DataResponsibly/DataSynthesizer/blob/master/notebooks/DataSynthesizer__correlated_attribute_mode.ipynb. Accesado el 15/10/2022.
11. Bohorquez, Nicolas. synthetic-data(DataSynthesizer). synthetic-data.ipynb. https://github.com/nickmancol/synthetic-data/blob/3184d35285147e005125b72c6cc65423e174151a/synthetic-data.ipynb. Accesado el 23/11/2022.
12. Python3. https://www.python.org/. Accesado el 3/08/2022

Capítulo 10 pp 145-172

FACTORES QUE INFLUYEN EN LA INNOVACIÓN DE LA CADENA LOGÍSTICA DE LAS EMPRESAS EN COLOMBIA ENTRE EL AÑO 2017 Y 2018

Diana Sofía Rondón Roa[1] y Linda Carolina Henao Rodríguez[1]

[1]Fundación Universitaria Konrad Lorenz

Resumen

La cadena logística ha tomado un papel fundamental en el comercio internacional, pues es la manera en que se conectan los proveedores con los clientes, a través de métodos de producción, distribución y entrega, a raíz de esto la innovación es un factor que ha tomado importancia, pues el mundo ha tenido cambios cada vez más radicales y así mismo debe acoplarse a las tendencias mundiales. Esta investigación busca determinar qué factores han motivado a las empresas a realizar innovaciones en métodos de producción, distribución, entrega, o sistemas logísticos en las empresas en Colombia en los años 2017 y 2018. Para tal fin se utilizó un árbol de decisión desarrollado a través de Machine Learning de aprendizaje supervisado. Como resultado de este estudio, se pudo determinar que los factores más relevantes para que una empresa implementará la innovación fue la adquisición de maquinaria y equipo y que perteneciera a actividades económicas específicas tales como el sector alimenticio, textil, químicos, plásticos y de metales. Es de suma importancia que se le dé prioridad a la innovación dentro de la cadena logística entendiendo que su adecuada gestión ayuda a potencializar y comercializar el producto haciendo que sea competitivo dentro del mercado global.

Introducción

La logística, definida por García et al. (2016) como una actividad interdisciplinaria que conecta varios procesos de la organización, desde la solicitud de insumos hasta el servicio postventa que se ofrece al cliente, engloba procesos como planificación, sistemas de producción, almacenamiento, embalaje de los productos y distribución, es un proceso fundamental en el funcionamiento de una empresa.

La evolución en los procesos logísticos es primordial para la adaptación de modelos de producción, distribución y comercialización modernos en una organización. Algunos sectores evolucionaron en la digitalización de sus procesos, aunque no es lo mismo que transformación digital, ayuda a la aceleración de la innovación en la economía del país (ANDI, 2016).

Por lo tanto, es importante mencionar el papel que juega la innovación en este contexto, esta se puede definir como la introducción de un nuevo, o significativamente mejorado producto (bien o servicio), de un proceso, de un nuevo método de comercialización o de un nuevo método organizativo, en las prácticas internas de la empresa, la organización del lugar de trabajo o las relaciones exteriores (OCDE, 2005). En este sentido, las tecnologías de la información y comunicación (TIC's) es el método que ayuda a lograr esta innovación que deben desarrollar todas las empresas en sus cadenas logísticas y así mismo ser competitivas. Según Estrada (2016), las TIC's y su función en la logística hace que factores como el tiempo y el espacio sean factores de subsistencia empresarial, más que factores de "valor agregado", y con ello el Internet de las Cosas y la Industria 4.0 hace que las transacciones entre cliente y empresa rompa las barreras que antiguamente eran casi imposibles de enfrentar sin estas herramientas, debido a esto, esta tecnología se ha convertido en un pilar fundamental para la cadena logística de toda empresa ya que incorpora servicios de control y gestión de altos volúmenes de información.

En Colombia, la Encuesta Nacional Logística (ENL) muestra el desempeño logístico que tuvo el país en comparación con el resto del mundo, midiéndose por componentes como el desempeño logístico, tercerización, comercio exterior, perspectivas de los servicios logísticos y competitividad regional. En el gráfico de comparación de componentes del costo logístico por actividades económicas para el año 2018 y 2020 que proporciona este informe, se evidencia que en el sector industria la categoría que más predomina en el 2018 es el transporte con un 46,2% seguido del almacenamiento con un 28,6%, en cambio para el 2020 la categoría más importante son los inventarios con 56,4% en segundo lugar el transporte con 20,1% (Departamento Nacional de Planeación, 2020).

En este contexto, surge la siguiente pregunta de investigación: "¿Cuáles son los factores que incidieron en las empresas colombianas para realizar innovaciones en el proceso logístico, tal como producción, distribución y comercialización entre el año 2017 y 2018?"

Para tal fin, en la primera parte del documento se analiza las definiciones de la logística y su importancia dentro del funcionamiento de una organización, además se presenta la revisión de la literatura e información de fuentes secundarias que permitió establecer una base teórica que identifique la importancia del estudio a desarrollar. En la segunda parte, se presenta la metodología que se aplicó en el presente estudio, en donde se utilizó un algoritmo llamado Machine Learning de aprendizaje supervisado, a saber, arboles de decisión, los cuales desarrollan sistemas de clasificación que predicen observaciones futuras con base en un conjunto de reglas o condiciones seleccionadas (IBM, 2021). Finalmente, se analizaron las variables seleccionadas por el algoritmo como factores que influyeron en la innovación, se presentaron las conclusiones del estudio y recomendaciones para futuras investigaciones.

Revisión de Literatura

La logística, actualmente, presenta una de las coyunturas más importantes dentro del comercio internacional, no solo por los nuevos retos que se han generado a partir de diversos factores, sino que su implementación añade un valor agregado a la competitividad que ha aumentado en los últimos años.

Como se mencionaba anteriormente en el documento, es una actividad interdisciplinaria que busca cumplir los objetivos de la compañía a través de diferentes procesos, es por ello, que esta área es la única transversal en la organización, pues está relacionada con unidades de negocio, compras, producción, distribución y procesos de entrega, lo que muestra que es una forma de generar valor y sobresalir en el sector estratégico y competitivo del mercado, por lo que su desarrollo debe ser considerado como una necesidad, no solo por las ganancias y dinero que puede generar, sino por el intercambio de información entre las diferentes partes que hacen parte de este proceso (Reyes et al. 2019).

Dentro del marco de la cadena logística es de fundamental importancia la comunicación entre las partes involucradas, se debe monitorear y colaborar entre cada una de las partes (proveedores y compradores) que logre el desarrollo de la cadena de suministro en un enfoque sustentable que logre el mejoramiento de los procesos como otro factor importante en la implementación de la innovación en las cadenas de suministro. La importancia de la implementación del enfoque monitoreado y colaborativo simultáneamente en el desarrollo del proveedor incrementa la producción ambienta. Evaluar la actuación de los proveedores puede ser guiada a través de los esfuerzos colaborativos como el entrenamiento, para identificar las áreas de mejoramiento.

Así mismo, la comunicación toma un rol determinante en el enfoque colaborativo para lograr el desarrollo sustentable, esta se ha determinado como uno de los primeros retos, entendiendo que puede acarrear conflictos y malentendidos en las partes. Sin embargo, esta también ayuda a la resolución de problemas entre compradores y vendedores. Una comunicación de manera asertiva entre los proveedores y compradores en todas las fases de la cadena de suministro garantiza la transmisión de la información pertinente a lo largo del proceso (Jadhav et al., 2019).

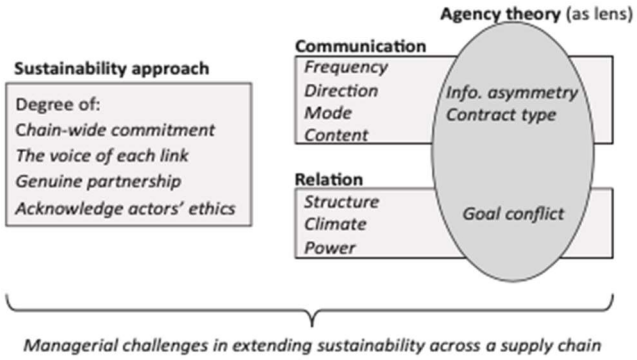


Figura 1 Comunicación en la cadena de suministro

Según un estudio titulado "Desafíos de la ampliación de la sostenibilidad en la cadena de suministro del transporte, el éxito de manejar la cadena de suministro digitalmente es uno de los desafíos estratégicos, si funciona y logra el éxito en la integración entre las partes de la cadena de suministro, como los proveedores, las empresas manufactureras y los clientes (Forslund et al., 2021).

En el estudio "*El emprendimiento digital y su impacto en las cadenas de suministro digitales: El papel mediador de las aplicaciones de inteligencia empresarial*" publicado por la Universidad Árabe de Tecnología en Jordania, en donde se buscaba la relación de la implementación de emprendimientos digitales dentro de las cadenas de suministro en los hoteles de Jordania y que implicaciones tenía, se logró identificar que las aplicaciones inteligencia de negocios afecta las cadenas de suministro digitales y por lo tanto incrementa el interés en este tipo de herramientas que podrían generar un impacto en el mejoramiento de las mismas (Awawdeh et al., 2021).

En el Índice de Desempeño Logístico (LPI, por sus siglas en inglés) se habla del desempeño logístico evaluado para los países de Latinoamérica generado por el Banco Mundial, que permite identificar oportunidades y retos en el desempeño de la logística comercial con 6 dimensiones logísticas de comercio, como lo son, la eficiencia del despacho de aduanas y gestión de fronteras, la calidad de la infraestructura relacionada con el comercio y el transporte, la facilidad de organizar envíos internacionales a precios competitivos, la competencia y calidad de los servicios logísticos, la capacidad de seguir y rastrear envíos, y la frecuencia con la que los envíos llegan a los destinatarios dentro del tiempo de entrega programado o esperado.

En el último reporte publicado por el Banco Mundial en 2018, permite dar una perspectiva de estado actual de los países de América Latina, entre ellos el mejor posicionado es Chile ocupando el puesto 34 a nivel mundial, seguido de México en el puesto 51 y Colombia en el puesto 68, de 165 países analizados.

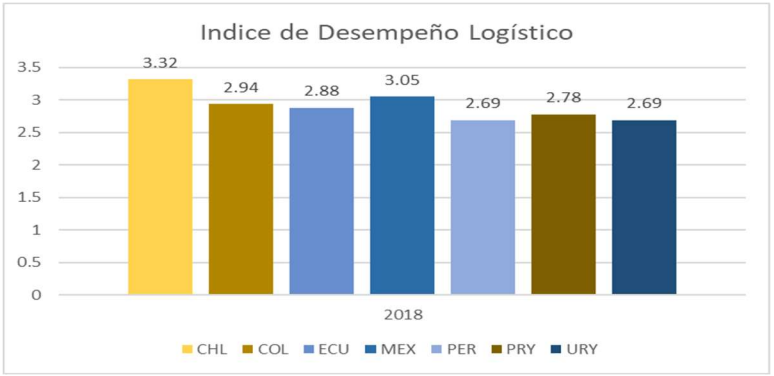


Figura 2 Elaboración propia a partir de datos del Banco Mundial

En países en vía de desarrollo como Colombia, no se cuenta con alto número de investigaciones que tenga como objetivo analizar la importancia de la generación de capacidades innovativas, y por esta razón tampoco con un perfil innovador que permita identificarlo en las organizaciones. En gran medida, se considera que esta tendencia se debe a que la tecnología que se usa actualmente proviene de países desarrollados que limitan el aprendizaje e investigación por conocimientos propios, impidiendo la creación de nuevas tecnologías y desarrollo de productos que se adapten al entorno actual (Robayo, 2016).

Cabe considerar que las PYMES que hacen parte del modelo económico de un país, debido a su tamaño y a la escasez de recursos económicos que manejan, es muy frecuente evidenciar que la persona que ejerce funciones gerenciales, realice actividades operacionales en las cuales no se encuentra capacitado, lo que conlleva a desarrollar actividades de bajo nivel por falta de conocimientos técnicos (Cano et al., 2014).

En segundo lugar, los textos relacionados con la importancia de la logística lograron identificar lo que significa este proceso dentro de una organización y la razón por la cual se debe dar una de las máximas prioridades para ser competitivo dentro del mercado. Según Manrique et al. (2019), las organizaciones deben desarrollar estructuras y ajustar procesos con los cuales puedan cumplir con las necesidades del cliente, sabiendo que este proceso es el resultado de los estándares de calidad que maneja una empresa, así mismo, este proceso alcanza el desarrollo y la potencialización de la producción y comercialización de los productos, conociendo cuales son los recursos con los que se cuenta, en que cantidad se requiere, y el aprovechamiento actual de ellos que permitirá a la compañía el alcance los objetivos que se ha propuesto.

Así mismo, Camacho et al. (2012) indica que la importancia de las áreas de la organización debe actuar juntas hacia un mismo objetivo y no independientemente, entendiendo que cada una tiene consecuencias en la otra. Estas interacciones deben dar relevancia al flujo de la empresa en el desarrollo del producto y la entrega final, con un sistema optimo, que no solo logre la comunicación interna en la empresa, sino que también tenga en cuenta el contexto externo en la que está involucrada.

La correcta integración de la cadena de suministro optimiza los recursos que comprende cada una de las áreas, estableciendo las oportunidades en proyectos y creación de estrategias que logren el cumplimiento de las metas en el mercado.

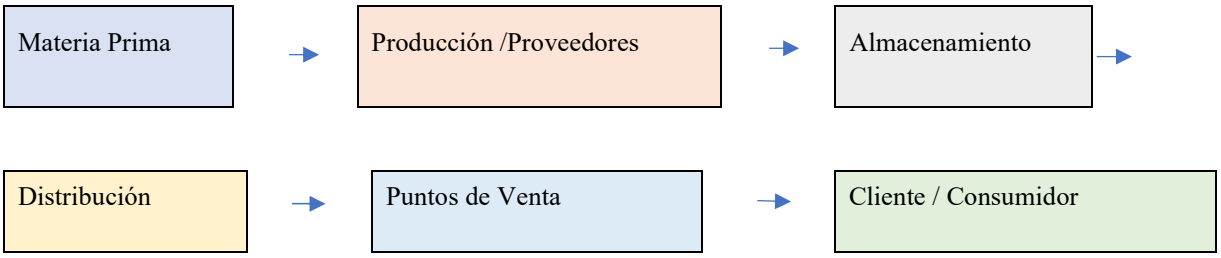


Figura 3 Elaboración propia
Nota: Flujo de la cadena de suministro

Por esta razón, la logística cumple un papel dentro de la competitividad, como herramienta estratégica, ya que, en la cadena logística, como adquisición de insumos, almacenaje y distribución de bienes y/o servicios, representa una oportunidad a la satisfacción del cliente y/o consumidor final, aprovechando el momento idóneo según el motivo y la necesidad del cliente (Martínez et al. 2018).

Para ejemplificar la innovación en el proceso de la cadena logística, Henríquez-Fuentes et al., (2018) realizó una medición de tiempos en las rutas de distribución en las que se busca planificar rutas optimas que permitan mejorar los tiempos y los cumplimientos en los horarios que se tienen programadas las entregas de los pedidos, como conclusión de este estudio se pudo implementar una metodología que permitía la planificación de las rutas con el fin de cumplir con los tiempos acordados y optimizando la capacidad del vehículo de entrega, pues en el análisis de esta herramienta se pudo evidenciar que se realizaba dos visitas a un mismo cliente en un mismo día lo que lleva a sobrecostos y reprocesos en la operación. Esta implementación permite evidenciar como se logra la innovación a través de procesos logísticos que ayuden a mejorar aspectos de la empresa.

En tercer lugar, la digitalización, la cual permite realizar cuatro operaciones básicas: generación y captación de información en varios formatos, ya sea audio, texto o video; transmisión, cómputo y almacenamiento los cuales se encuentran interconectados simultáneamente componiendo el sistema de tecnologías de la información y las comunicaciones (TIC's) (Jordan et al. 2013), que logra potencializar tendencias como la Industria 4.0 a través del Internet de las Cosas y los sistemas ciber físicos (sistemas inteligentes y autónomos que utilizan Inteligencia Artificial) para controlar activos tangibles, tales como, maquinas robots y vehículos. Esto hace que la logística sea más inteligente, desde la fabricación hasta el almacenamiento, así mismo, se interconecta con otros sistemas o softwares para tener una mayor visibilidad y control, formando un papel fundamental en la transformación digital de cualquier empresa.

Henríquez (2020) muestra un panorama general de uso de tecnologías incluidas en este proceso, resaltando la importancia que ha tenido desde la pandemia COVID-19 en las PYMES (pequeñas y medianas empresas), indicando que el 90% del total de estas organizaciones tienen internet, pero solo lo usan para enviar y recibir correos electrónicos, un 40% tiene página web y el 27% realiza comercio electrónico. Este factor que antes se veía como un complemento opcional, ahora se ha vuelto requisito para la competitividad en el mercado, comenzando a utilizar herramientas como dispositivos tecnológicos (computadores, servidores, etc.), softwares de ventas, marketing y gestión de clientes adaptado a las necesidades de la compañía y sistemas de ciberseguridad, entendiendo que las PYMES dirigen sus recursos humanos y financieros para las eventualidades que se presentan al día, es por ello que requieren apoyos para el financiamiento y asesoramiento técnico para la implementación de estas nuevas estrategias.

Según la CEPAL (2019), el futuro de la cadena logística esta direccionada a la interconectividad de la información, la optimización de tiempos y recursos, con una inversión considerable en el desarrollo de la innovación para ser competitivos en el mercado, pues el valor agregado de cada bien o servicio está siendo fuertemente influenciado por el intercambio de información de datos a través de herramientas tecnológicas entre los participantes de la logística. Adicionalmente, herramientas como las aplicaciones, softwares o sistemas informáticos combinan y coordinan diferentes tecnologías de control, transferencia y procesamiento de información; con tal de mejorar la eficiencia, seguridad y sostenibilidad de los servicios en la infraestructura de cada compañía. Para lograr este propósito, estas tecnologías recolectan, procesan y transmiten información de transacciones comerciales, de las diferentes operaciones de la mercancía, tráfico y otras variables que influyen en el flujo de bienes, generando oportunidades para generar valor agregado que diferencien de la competencia y una disminución en los factores externos negativos tanto sociales como ambientales.

Se empieza a hablar de cadena de suministro 4.0 cuando se integran herramientas tecnológicas en los flujos de información del proceso, en donde la información debe ser fiable y en tiempo real, y que así mismo ayuda a la toma de decisiones de una manera más asertiva, individual y consistente. Esto también aportando a que los procesos operativos sean automatizados, optimizando su ejecución y que sean rápidos, eficaces y eficientes (Budet Jofra et al., 2018).

Aunque en este sentido, América Latina y el Caribe presenta varios desafíos para lograr esta integración tecnológica, como por ejemplo la brecha digital que existe con países desarrollados, pues tienen acceso a tecnología más avanzada y cooperación entre el uso de estas; ausencia de normativas que incluyan de manera adecuada la ciberseguridad, uso

de patentes y regulación que ayuden a la cooperación internacional; la falta de inversión en investigación y desarrollo; y la falta de apoyo gubernamental que fomente capacitaciones tecnológicas que haga más productiva y eficiente la logística, y así mismo, el comercio internacional en toda la región (CEPAL, 2019).

Por esta razón, es necesario que el conocimiento en innovación reciba una inversión decidida, que fomente la adopción de nuevas y mejoradas prácticas, que no solo requiere aprendizaje de distintas teorías, sino que las personas que lo implementan tengan una actitud abierta, activa y dinámica, que ayude a generar valor agregado a la gestión de procesos (De la Hoz Granadillo et al., 2017).

Adicional por lo mencionado anteriormente, un estudio realizado por IBM (Institute for Business Value) en cooperación con Oxfords Economics entrevistó a 1500 directores de cadenas de suministro (CSCOs, por sus siglas en ingles) de más de 35 países pertenecientes a 24 industrias, en donde se buscaba identificar sus perspectivas respecto al liderazgo del negocio, sus desafíos y responsabilidades dentro del campo, la sostenibilidad, incluyendo como dirigen y manejan estos desafíos y que oportunidades pueden ver para el futuro (IBM Institute for Business Value , 2022).

Dentro de esta investigación se logró identificar que más del 47% de los CSCOs han introducido nuevas tecnologías de automatización con el fin de añadir predictibilidad, flexibilidad e inteligencia en las operaciones para la toma de decisiones. En la tabla 1 se puede observar que soluciones han adoptado y en qué porcentaje: un poco menos que la mitad de los entrevistados informaron que rebalancearon la fuerza de trabajo, seguido por la colaboración con proveedores para manejar inconvenientes, adoptaron nuevas tecnologías incluyendo la automatización, rebalancearon el inventario hacia los clientes con un 42% y por último reasignaron personal en funciones alternas que permita ser mucho más eficiente su horario laboral.

Soluciones	Porcentaje de implementación
Rebalanceo de fuerza de Trabajo	48%
Colaboración con proveedores para manejar inconvenientes	48%
Adopción de nuevas tecnologías, incluyendo la automatización	47%
Rebalanceo de inventario a clientes	42%
Personal reasignado a funciones alternativas	41%

Figure 3 Elaboración propia a partir de datos del IBM

Algunos de los directores de las cadenas de suministro han utilizado una serie de herramientas de Inteligencia Artificial (AI por sus siglas en inglés) con el fin de soportar la base de datos para la toma de decisiones, ayudar a las cadenas de suministro a identificar con mayor rapidez, a priorizar y recomendar las próximas decisiones.

El estudio indica, que la automatización también ayuda a las habilidades de los empleados, pues ayuda a reducir el desperdicio de los procesamientos cognitivos de análisis similares, decisiones y acciones, lo cual los puede liberar para enfocarse en responsabilidades estratégicas, analíticas y de valor agregado. Con una mirada más profunda de la implementación de IA aumenta los puntos fuertes, complementa sus debilidades y empodera su organización para enfocarse en lo que es realmente importante. Esto ayuda a que se destaque lo mejor de los empleados y la tecnología, creando un mayor valor a través de la cadena de suministro.

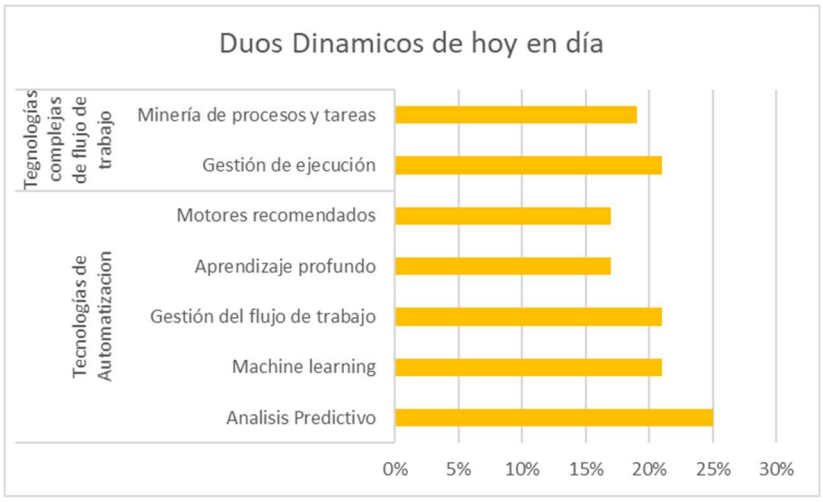


Figura 4 Elaboración propia a partir de datos del IBM

Nota: Las nuevas tecnologías son más poderosas a través de las cadenas de suministro de las organizaciones.

En la figura 6, se puede observar cómo los CSCOs priorizan las tecnologías hablando de su entusiasmo para los flujos inteligentes, más del 54 % se enfoca en la inteligencia artificial y el machine learning y un poco menos con el 49% se enfoca en las nubes hibridas y el internet de las cosas (IoT por sus siglas en ingles). Estos lideres no se detienen al momento de implementar flujos de inteligencia internos para su organización, se espera que el 32% de estos flujos estén implementados dentro del ecosistema con otros proveedores para el 2030.

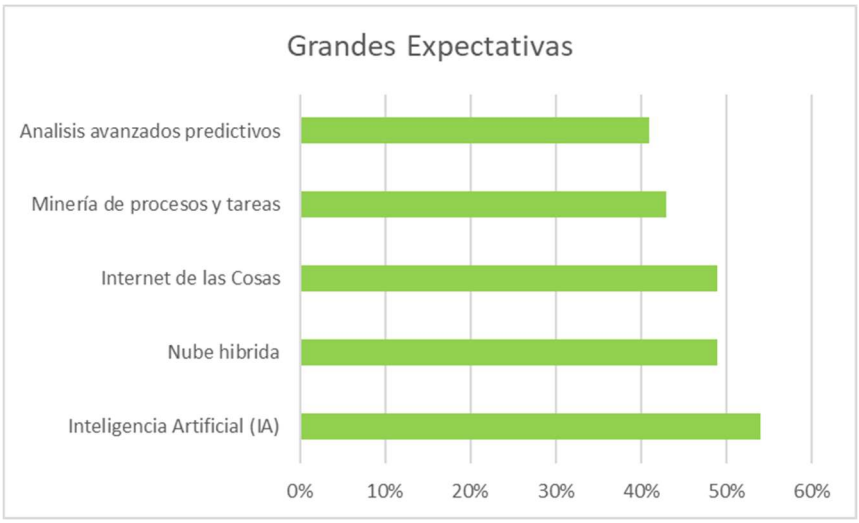


Figura 5 Elaboración propia a partir de datos del IBM

Nota: La mayoría de los CSCOs creen que la Inteligencia Artificial dará resultados en los próximos 3 años.

El uso de estas herramientas tecnológicas es mucho más frecuente y requerirán de inversiones de diferentes partes de la cadena logística, no solo en la adquisición de tecnología sino también en la redefinición de procesos con las que se interactuarán con la gestión de la compañía y organización. Con esto se busca añadir valor a la experiencia del cliente, mediante servicios personalizados y tiempos de entrega más eficientes que se adecuen a las necesidades del mercado. Algunas herramientas que sobresalen son el Blockchain, Inteligencia Artificial, internet de las cosas (IoT), sistemas recomendadores y entregas automatizadas.

Por último, en la categoría de fomento de programas gubernamentales que impulsan el uso de herramientas tecnológicas, la Secretaría General Iberoamericana (2021) comprobó que la digitalización más que el futuro, es el presente de las empresas en donde deben juntar esfuerzos el sector público con el sector privado para fomentar e impulsar el desarrollo en las sociedades. Realizan un importante énfasis en la educación, la formación y la capacitación, la cual debe ser inclusiva, igualitaria y accesible a todos los sectores, como ejemplo, varios delegados de los países de Republica Dominicana, Ecuador, México y Chile expusieron las medidas que actualmente están promoviendo en cada uno de sus Estados para el fomento de la transformación digital. En primera medida, República Dominicana procedió a la regulación del teletrabajo, políticas públicas para proporcionar asistencia técnica en la digitalización, fomento de capacitación al acceso del conocimiento de la política financiera. Adicional, puso en marcha el Programa de Encadenamiento Productivo Digital en la que las empresas tractoras se les brinda un subsidio para el diseño de programas que involucren PYMES y emprendimientos en herramientas digitales, y firmo un memorándum con el gobierno de España para compartir nuevas prácticas y aprender sobre el modelo español relacionado con la industria 4.0, la innovación y la digitalización. En segundo lugar, Ecuador puso en marcha la Estrategia Nacional de Comercio Electrónico, basada en tres pilares: el fomento del e-commerce para las mipymes, mejorar los sistemas de pago y avanzar en el ámbito de la logística, además, desarrollaron diversas acciones para reforzar la capacidad de las pequeñas y medianas empresas no solo en temas de bancarización, sino en el fortalecimiento de habilidades blandas, con organización de charlas y webinars. Por otro lado, México implemento estrategias de comercio electrónico, llevando a cabo iniciativas para conectar cadenas de valor de manera digital y puso en marcha el Programa 4.0 que es un programa de la sociedad para la sociedad en donde participan diferentes actores tanto el gobierno como la industria, con el objetivo de crear una sociedad digital a través del trabajo en la ciudadanía fomentando el talento que requieren las nuevas tecnologías. Y, por último Chile puso en marcha el programa "Ruta Digital" desde SERCOTEC (Servicio de Cooperación Técnica de Chile) en donde promueve la capacitación y la transformación desde el cambio de la mentalidad, donde los empresarios pueden diagnosticar el nivel de digitalización en el que se encuentran y así acceder a cursos en materia de ciberseguridad, manejo de inventario, marketing o ventas, así mismo, el programa brinda un fondo de financiamiento, basado en un kit tecnológico que permite al usuario poner en práctica lo aprendido en los diferentes cursos. De igual manera, el gobierno nacional de Colombia está avanzando el país en materia de trasformación digital, mostrando iniciativas implementadas como la adopción de la nube por parte de las entidades públicas, interoperabilidad en 100 entidades, Plan Nacional de Talento en Habilidades Digitales, aceleración y fortalecimiento de servicios digitales en el sector público; y la creación de la oficina de inteligencia artificial (Muñoz et al., 2020).

Metodología

La metodología es cuantitativa, pues busca que el resultado se obtenga de datos cuantitativos y medibles, y de esta manera, de corte transversal pues se toman datos de un solo periodo de tiempo. Para este estudio se usó un algoritmo Machine Learning de aprendizaje supervisado llamado Arboles de Decisión de Clasificación para una variable binaria, con su función principal de clasificar los datos seleccionados y analizados a través de las herramientas del software JAPS versión 0.16.3 para Windows.

Esta metodología consiste en el Manual de Python, entendiendo que este es un lenguaje de programación potente y de fácil aprendizaje que cuenta con estructuras de información de alto nivel con un eficiente sistema de programación enfocado en los objetos. Este lenguaje es ideal para scripting y desarrollo eficaz de aplicaciones (Python Software Foundation, 2022).

Los ajustes algorítmicos de este estudio se configuraron de la siguiente manera: el número mínimo de observaciones por rama fueron 20, el número de observaciones en un nuevo terminal es fue 7 y el máximo de profundidad de los nodos o interacciones fueron de 30 para el árbol final. Las variables se escalaron porque estaban en diferentes unidades de medida y se sembró una semilla aleatoria para replicar los resultados la cual fue "123".

El árbol de decisiones de clasificación como se dijo anteriormente es de aprendizaje supervisado que utiliza un árbol de decisión como un modelo predictivo que va de las observaciones de un dato (que se representa en la raíz de este) a la conclusión sobre el valor objetivo del dato (representado por los puntos finales del árbol).

En los ajustes del árbol de clasificación, no se puede usar RSS (suma de los residuos cuadrados) como criterio para cortes binarios, como alternativa se utiliza la tasa de error de clasificación que es una simple fracción de las observaciones de entrenamiento en la región que no pertenece a la clase más común:

$$E = 1 - \frac{\max}{k} (\hat{P}mk)$$

En esta ecuación el $\hat{P}mk$ representa la proporción de observaciones de entrenamiento en la región *m*th región que de la clase *k*th. Sin embargo, el error de clasificación no es lo suficientemente sensible para el crecimiento del árbol, y en la práctica otras 2 medidas son preferibles.

El índice de Gini es definido de la siguiente manera:

$$G = \sum_{k=1}^{K} \hat{P}\, mk\, (1 - \hat{P}mk)$$

Una medida total de la varianza total a través de las clases K. No es muy difícil de ver que el índice de Gini toma un valor pequeño si todos los $\hat{P}mk$ son cercanos a cero o a uno. Por esta razón el índice de Gini es referenciado como una medida pura de los nodos (intersecciones), un pequeño valor indica que el nodo contiene observaciones predominantes de una clase singular.

Una alternativa a este indicador es la entropía cruzada dada por:

$$D = \sum_{k=1}^{K} \hat{P}\, mk\ \log \hat{P}\, mk$$

Desde que 0 £ $\hat{P}mk$ £ 1, se deduce que 0 £ $\hat{P}mk \log \hat{P}\, mk$. Una vez que muestra que entropía cruzada tomará un valor cercano a 0 si $\hat{P}mk$ son todos muy cercanos a 0 o a 1. Por esta razón, similar al índice de Gini, la entropía cruzada tomará un valor pequeño, de igual manera, estos dos indicadores resultan bastante similares numéricamente.

Al crear un árbol de clasificación, el índice de Gini o la entropía cruzada se utilizan para evaluar la calidad de un determinado corte, pues estos son enfoques son muchos más sensibles a la pureza de las intersecciones que la tasa de error de la clasificación. Se recomienda que se utilice la tasa de error de clasificación cuando el objetivo de la precisión es la predicción del árbol podado al final (James et al., 2013).

Datos

La presente investigación tomó los datos de la "Encuesta de Desarrollo e Innovación Tecnológica - EDIT - Industria - 2017 – 2018" desarrollada por el Departamento Administrativo Nacional de Estadística (DANE), en su versión más actualizada publicada en el año 2020 correspondiente a los años 2017 y 2018. Según la ficha técnica de la encuesta, publicada el 8 de enero de 2020, para la selección de la población y el muestreo de la aplicación del estudio, se segregaron por empresas industriales que tienen en su establecimiento de 10 o más personas empleadas o, por el contrario, que registren un valor de producción al año igual o mayor a $506 millones de pesos colombianos, correspondiente al directorio de empresas de la Encuesta Anual Manufacturera (EAM).

Según el parámetro establecido, el estudio muestra información para 55 subsectores o actividades industriales de acuerdo con la Clasificación Internacional Industrial Uniforme (CIIU), aplicadas a Colombia, con la aplicación de 635 variables las cuales resumen de forma estratégica y metodológica implementadas por la Organización de Cooperación y Desarrollo Económico (OCDE) y el Manual de Oslo. Se adaptaron las recomendaciones de acuerdo con las necesidades identificadas para Colombia.

Entre los subsectores seleccionados, se encuentran el procesamiento y conservación de carnes y pescado; frutas, legumbres, hortalizas y tubérculos; elaboración de aceites, productos lácteos, azúcar y bebidas; procesos industriales de textiles, prendas de vestir y calzado; fabricación de productos de madera; productos petroleros, fabricación de plásticos y cauchos, entre otros.

Este estudio tiene cobertura a nivel nacional dentro de cada una de las variables definidas.

La recolección de datos se realizó con un inicio y final en el año 2019 con un ciclo bienal, esta se realizó con un auto diligenciamiento de formulario electrónico vía internet (DANE, 2020).

Variables

La variable dependiente es la respuesta a la siguiente pregunta "Indique si durante el periodo 2017-2018 su empresa introdujo nuevos o significativamente mejorados métodos de producción, distribución, entrega, o sistemas logísticos en su empresa", en cuyas opciones de respuesta se debía marcar SI=1 o NO=2.

Dentro de las variables dependientes, se seleccionaron las preguntas que más relevancia e importancia obtuvieron dentro la encuesta escogida como medio de estudio para lograr el objetivo general propuesto para esta investigación. Es por ello, que a continuación se relaciona la tabla 1, en donde se detalla la notación (código de la pregunta), variable (aspecto a medir), tipo de variable y pregunta.

Tabla 1 Variables seleccionadas para el estudio.

NOTACIÓN	VARIABLE	TIPO DE VARIABLE	PREGUNTA
I3R1C1	Ventas nacionales totales 2017 (Miles de pesos corrientes)	Continua	"Indique el valor correspondiente a los ingresos o ventas operacionales nacionales y las exportaciones efectuadas por su empresa en los años 2017 y 2018. (En miles de pesos Corrientes)"
I3R1C2	Exportaciones totales 2017 (Miles de pesos corrientes)	Continua	"Indique el valor correspondiente a los ingresos o ventas operacionales nacionales y las exportaciones efectuadas por su empresa en los años 2017 y 2018. (En miles de pesos corrientes)"
I3R2C1	Ventas nacionales totales 2018 (Miles de pesos corrientes)	Continua	"Indique el valor correspondiente a los ingresos o ventas operacionales nacionales y las exportaciones efectuadas por su empresa en los años 2017 y 2018. (En miles de pesos corrientes)"
I3R2C2	Exportaciones totales 2018 (Miles de pesos corrientes)	Continua	"Indique el valor correspondiente a los ingresos o ventas operacionales nacionales y las exportaciones efectuadas por su empresa en los

			años 2017 y 2018. (En miles de pesos corrientes)"
II1R1C1	Actividades de I+D internas. Monto invertido 2017	Continua	"Indique el valor invertido por su empresa en los años 2017 y 2018, en cada una de las siguientes actividades científicas, tecnológicas y de innovación, para la introducción de bienes o servicios nuevos o significativamente mejorados, y/o la implementación de procesos nuevos o significativamente mejorados, de métodos organizativos nuevos, o de técnicas de comercialización nuevas."
II1R1C2	Actividades de I+D internas. Monto invertido 2018	Continua	"Indique el valor invertido por su empresa en los años 2017 y 2018, en cada una de las siguientes actividades científicas, tecnológicas y de innovación, para la introducción de bienes o servicios nuevos o significativamente mejorados, y/o la implementación de procesos nuevos o significativamente mejorados, de métodos organizativos nuevos, o de técnicas de comercialización nuevas."
II1R2C1	Adquisición de I+D (externa). Monto invertido 2017	Continua	"Indique el valor invertido por su empresa en los años 2017 y 2018, en cada una de las siguientes actividades científicas, tecnológicas y de innovación, para la introducción de bienes o servicios nuevos o significativamente mejorados, y/o la implementación de procesos nuevos o significativamente mejorados, de métodos organizativos nuevos, o de técnicas de comercialización nuevas."
II1R2C2	Adquisición de I+D (externa). Monto invertido 2018	Continua	"Indique el valor invertido por su empresa en los años 2017 y 2018, en cada una de las siguientes actividades científicas, tecnológicas y de innovación, para la introducción de bienes o servicios nuevos o significativamente mejorados, y/o la implementación de procesos nuevos o significativamente mejorados, de métodos organizativos nuevos, o de técnicas de comercialización nuevas."
II1R3C1	Adquisición de maquinaria y equipo. Monto invertido 2017	Continua	"Indique el valor invertido por su empresa en los años 2017 y 2018, en cada una de las siguientes actividades científicas, tecnológicas y de innovación, para la introducción de bienes o servicios nuevos o significativamente mejorados, y/o la implementación de procesos nuevos o significativamente mejorados, de métodos organizativos nuevos, o de técnicas de comercialización nuevas."

II1R3C2	Adquisición de maquinaria y equipo. Monto invertido 2018	Continua	"Indique el valor invertido por su empresa en los años 2017 y 2018, en cada una de las siguientes actividades científicas, tecnológicas y de innovación, para la introducción de bienes o servicios nuevos o significativamente mejorados, y/o la implementación de procesos nuevos o significativamente mejorados, de métodos organizativos nuevos, o de técnicas de comercialización nuevas."
IV1R1C1	Doctorado. Personal ocupado promedio 2017	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R1C2	Doctorado. Personal ocupado promedio 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R1C3	Doctorado. Personal ocupado promedio que participó en la realización de ACTI 2017	Discreta	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R1C4	Doctorado. Personal ocupado promedio que participó en la realización de ACTI 2018	Discreta	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R2C1	Maestría. Personal ocupado promedio 2017	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R2C2	Maestría. Personal ocupado promedio 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."

IV1R2C3	Maestría. Personal ocupado promedio que participó en la realización de ACTI 2017	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R2C4	Maestría. Personal ocupado promedio que participó en la realización de ACTI 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R3C1	Especialización. Personal ocupado promedio 2017	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R3C2	Especialización. Personal ocupado promedio 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R3C3	Especialización. Personal ocupado promedio que participó en la realización de ACTI 2017	Discreta	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R3C4	Especialización. Personal ocupado promedio que participó en la realización de ACTI 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R4C1	Universitario. Personal ocupado promedio 2017	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R4C2	Universitario. Personal ocupado promedio 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y

			2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”
IV1R4C3	Universitario. Personal ocupado promedio que participó en la realización de ACTI 2017	Continua	“Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”
IV1R4C4	Universitario. Personal ocupado promedio que participó en la realización de ACTI 2018	Continua	“Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”
IV1R5C1	Tecnólogo. Personal ocupado promedio 2017	Continua	“Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”
IV1R5C2	Tecnólogo. Personal ocupado promedio 2018	Continua	“Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”
IV1R5C3	Tecnólogo. Personal ocupado promedio que participó en la realización de ACTI 2017	Continua	“Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”
IV1R5C4	Tecnólogo. Personal ocupado promedio que participó en la realización de ACTI 2018	Continua	“Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”
IV1R6C1	Técnico profesional. Personal ocupado promedio 2017	Continua	“Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido.”

IV1R6C2	Técnico profesional. Personal ocupado promedio 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R6C3	Técnico profesional. Personal ocupado promedio que participó en la realización de ACTI 2017	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."
IV1R6C4	Técnico profesional. Personal ocupado promedio que participó en la realización de ACTI 2018	Continua	"Indique el personal ocupado promedio que laboró en su empresa en los años 2017 y 2018. De éste, especifique el número que participó en la realización de actividades científicas, tecnológicas y de innovación en los años 2017 y 2018, de acuerdo con el máximo nivel educativo alcanzado y con título obtenido."

Nota. Elaboración propia a partir de información de la encuesta seleccionada para el estudio.

Resultados

Los hallazgos que se obtuvieron después del estudio y la metodología aplicada a este estudio fueron los siguientes:

- Data Split

Es un algoritmo en el que se trabajaron con una muestra de 1850 empresas, de las cuales se usaron 370 para validación y 1480 para entrenamiento.

- Proporción de clases

Tabla 2 Proporción de Clases

Class Proportions

	Data Set	Training Set	Test Set
0	0.484	0.477	0.511
SI	0.516	0.523	0.489

Fuente: Elaboración propia usando Japs.

Ayuda a mostrar la proporción de cada conjunto mencionados en la ilustración anterior, la cantidad de datos proporcional al total, uno para validación y otro para entrenamiento. Se puede concluir que las empresas que no innovaron eran el 48% del total de las empresas evaluadas, y para el entrenamiento se usó el 47% del 100% de las empresas no innovaron y el 51% para la parte de validación.

- Métricas de Evaluación

Tabla 3 Evaluación de Métricas.

Evaluation Metrics

	0	SI	Average / Total
Support	189	181	370
Accuracy	0.700	0.700	0.700
Precision (Positive Predictive Value)	0.701	0.699	0.700
Recall (True Positive Rate)	0.720	0.680	0.700
False Positive Rate	0.320	0.280	0.300
False Discovery Rate	0.299	0.301	0.300
F1 Score	0.710	0.689	0.700
Matthews Correlation Coefficient	0.400	0.400	0.400
Area Under Curve (AUC)	0.700	0.700	0.700
Negative Predictive Value	0.699	0.701	0.700
True Negative Rate	0.680	0.720	0.700
False Negative Rate	0.280	0.320	0.300
False Omission Rate	0.301	0.299	0.300
Threat Score	0.805	0.750	0.777
Statistical Parity	0.524	0.476	1.000

Note. All metrics are calculated for every class against all other classes.

Indica las mediciones más usadas comúnmente en las métricas de evaluación de un estudio como, por ejemplo, la precisión indica la proporción de los datos correctos dentro de los datos seleccionados, que para este caso fue 0.70, es un resultado aceptable teniendo en cuenta la cantidad de muestras utilizadas para el estudio. También se encuentra, la medición Recall (True Positive Rate) (Recuperación (tasa de verdaderos positivos)), la cual indica los datos previstos como positivos, dentro de los datos totales positivos (Microsoft, 2022)

- Matriz de Importancia Relativa

Tabla 4 Matriz de Importancia Relativa.

Feature Importance

	Relative Importance
Adquisición de maquinaria y equipo. Monto invertido 2018	24.155
Adquisición de maquinaria y equipo. Monto invertido 2017	23.126
CIIU4	18.220
Ventas nacionales totales 2017 (Miles de pesos corrientes)	7.949
Universitario. Personal ocupado promedio que participó en la realización de ACTI 2017	7.671
Ventas nacionales totales 2018 (Miles de pesos corrientes)	7.437
Universitario. Personal ocupado promedio 2018	6.979
Exportaciones totales 2018 (Miles de pesos corrientes)	0.858
Tecnólogo. Personal ocupado promedio que participó en la realización de ACTI 2017	0.547
Exportaciones totales 2017 (Miles de pesos corrientes)	0.513
Universitario. Personal ocupado promedio 2017	0.478

Feature Importance

	Relative Importance
Actividades de I+D internas. Monto invertido 2017	0.444
Actividades de I+D internas. Monto invertido 2018	0.444
Tecnólogo. Personal ocupado promedio que participó en la realización de ACTI 2018	0.368
Técnico profesional. Personal ocupado promedio que participó en la realización de ACTI 2018	0.263
Doctorado. Personal ocupado promedio que participó en la realización de ACTI 2017	0.229
Universitario. Personal ocupado promedio que participó en la realización de ACTI 2018	0.202
Doctorado. Personal ocupado promedio que participó en la realización de ACTI 2018	0.115

- Splits in Tree

Tabla 5 Cortes de cada rama del árbol.

Splits in Tree

	Obs. in Split	Split Point	Improvement
Adquisición de maquinaria y equipo. Monto invertido 2017	1480	-0.222	66.469
Adquisición de maquinaria y equipo. Monto invertido 2018	830	-0.213	48.578
CIIU4	297	4.000	15.608
CIIU4	650	5.000	23.382
CIIU4	223	6.000	9.616

Note. For each level of the tree, only the split with the highest improvement in deviance is shown.

En este se muestra los puntos de corte que realizó el programa para realizar la clasificación, se recuerda que se realizó una estandarización de las variables por lo que estos tipos de corte no corresponden a los valores originales.

Para identificar el valor de inversión que se realizó en cada uno de los puntos de corte seleccionados por el software, se identificó la ecuación que está programada:

$$z = \frac{(n - \bar{x})}{\sigma}$$

Y se despejó para identificar la variable *n*, dando el siguiente resultado:

$$(z * \sigma) + \bar{x} = n$$

Como resultado, arrojo los siguientes valores para las variables II1R3C1 (Adquisición de maquinaria y equipo. Monto invertido 2017) y II1R3C2 (Adquisición de maquinaria y equipo. Monto invertido 2018):

Tabla 6 Resultados despejo ecuación de estandarización.

	II1R3C1	II1R3C2
Media	464068.18	442231.5
Desviación Estándar	20390000	20780000
Valor (en millones)	15488.177	-382.55

Según los anexos, los datos recolectados están en miles de pesos, es decir que para la variable II1R3C1 el punto de corte lo realizó en las empresas que invirtieron más de $15’488.177 de pesos y para la variable II1R3C2 las que invirtieron menos de $382.550.

- Curva en ROC

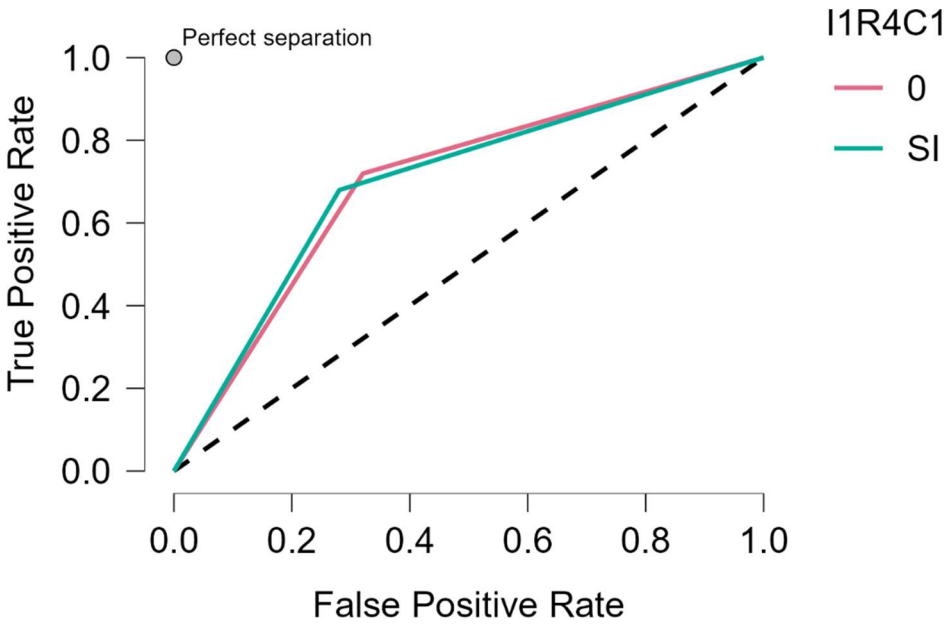


Figura 6 Curva ROC

Como se puede ver la curva de ROC no tiene un margen de precisión aceptable dada el tamaño de la muestra que es bastante grande y también se puede observar que cumple con los estándares porque predice de la misma manera para SI y para NO, además que muestra que no está sobre la línea a 45° que mostraría que el modelo no tiene la capacidad de predecir porque no podría distinguir entre el SI y el NO.

- Andrews Curves Plot

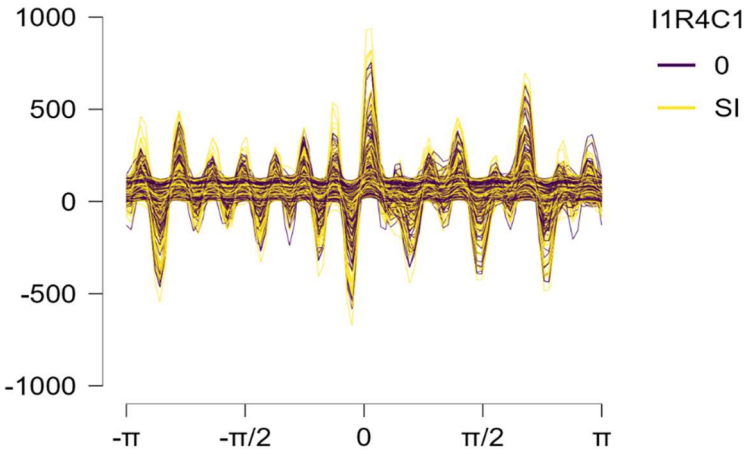


Figura 7 Curva de Andrews

Este grafico está basado en la visualización de datos multivariantes en dos dimensiones, su magnitud afecta la frecuencia, la amplitud y la periodicidad de las variables (Schiattino & Silva, 2013).

- Decisión Tree Plot

Como resultado de la aplicación de este modelo, se generaron 3 nodos terminales que cumplieron con las condiciones dadas, uno tiene 150 empresas, otro 196 y otro 427.

Entre estos, el primer nodo terminal (150 empresas) cumplía con las siguientes condiciones: que la primera variable II1R3C1 (Adquisición de maquinaria y equipo. Monto invertido 2017) tuviera una inversión en innovación menor a -0.22, que con la estandarización realizada, dio un valor monetario de $15’488.177; la segunda condición era la variable II1R3C2 (Adquisición de maquinaria y equipo. Monto invertido 2018) tuviera un valor invertido mayor a -0,213, es decir, una valor menor a $382,550; y la última condición que perteneciera a una de las 23 CIIU, tales como, procesamiento y conservación de carne y productos cárnicos; elaboración de productos de molinería, productos de panadería, productos alimenticios (Cacao, chocolate y productos de confitería), macarrones, fideos, alcuzcuz y productos farináceos similares, alimentos preparados para animales; destilación, rectificación y mezcla de bebidas alcohólicas; elaboración de bebidas no alcohólicas, producción de aguas minerales y de otras aguas embotelladas, fabricación de otros productos textiles (Confección de artículos con materiales textiles, excepto prendas de vestir), fabricación de recipientes de madera, entre otros (Anexo 1).

Para los siguientes nodos terminales debían cumplir con dos condiciones: que para la primera variable II1R3C1 (Adquisición de maquinaria y equipo. Monto invertido 2017) tuviera un valor monetario invertido en innovación mayor a -0.22, es decir, $15’488.177; y la segunda condición que perteneciera a dos grupos de CIIU, por un nodo terminal (196 empresas) que perteneciera a 36 CIIU, entre las que destacan: procesamiento y conservación de carne y productos cárnicos, elaboración de productos lácteos, productos de molinería, refinación de azúcar, productos de panadería, alimentos preparados para animales, bebidas no alcohólicas, producción de aguas minerales y de otras aguas embotelladas; destilación, rectificación y mezcla de bebidas alcohólicas; preparación e hilatura de fibras textiles; acabado de productos textiles; confección de prendas de vestir, excepto prendas de piel, fabricación de calzado de cuero y piel, con cualquier tipo de suela, otros tipos de calzado, excepto calzado de cuero y piel, partes del calzado; aserrado, acepillado e impregnación de la madera, entre otros. (Anexo 2). Finalmente para en el tercer nodo terminal (427 empresas) que perteneciera a 34 CIIU, entre las cuales están procesamiento y conservación de pescados, crustáceos y moluscos, elaboración de almidones y productos derivados del almidón; productos de café (Trilla de Café), productos de café (Descafeinado, tostión y molienda del café), otros productos alimenticios (Cacao, chocolate y productos de confitería), fabricación de productos textiles (tejeduría de productos textiles), otros productos textiles (tejidos de punto y ganchillo), otros productos textiles (Confección de artículos con materiales textiles, excepto prendas de vestir), fabricación de sustancias y productos químicos básicos, plásticos en formas primarias, pinturas, barnices y revestimientos similares, tintas para impresión y masillas, etc. (Anexo 3).

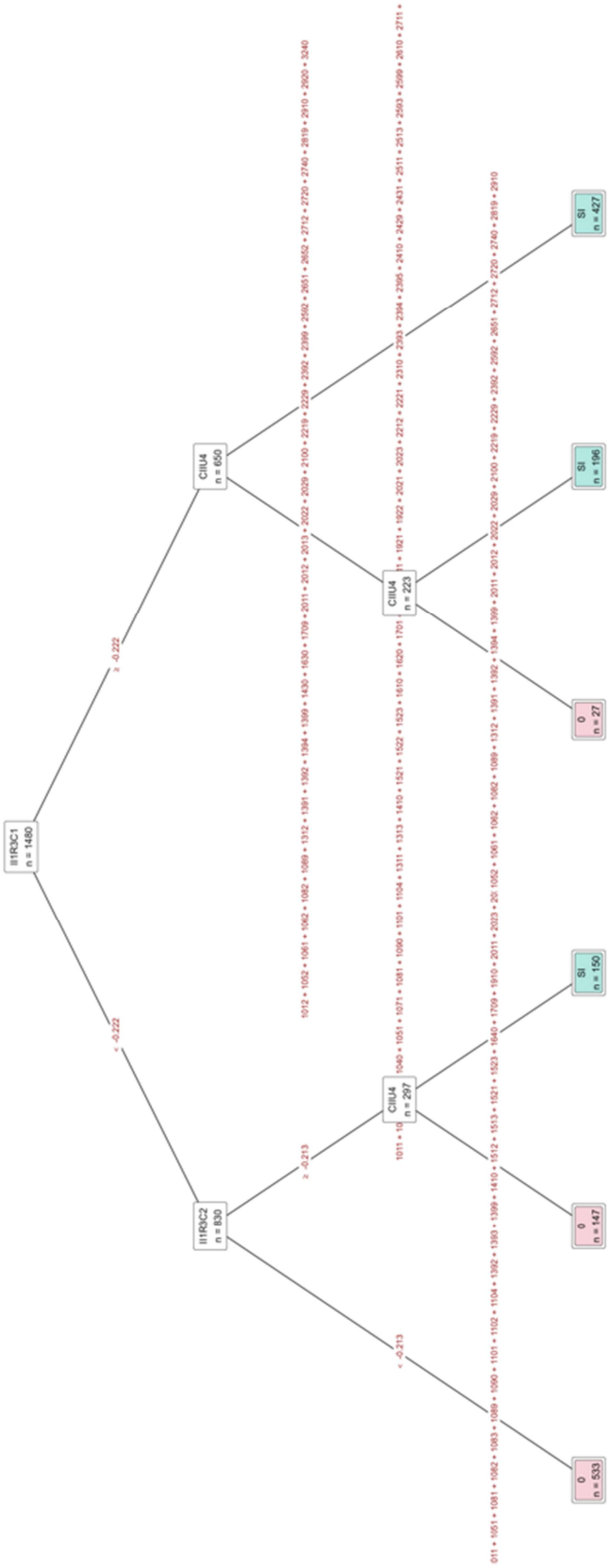


Ilustración 1 Árbol de decisión generado a partir de los resultados ingresados.

Discusión y Conclusiones

En este artículo se presentó un modelo de machine-learning que pretendía identificar las variables que eran relevantes para determinar los factores que más influían para que una empresa pudiera innovar en la industria, en el cual se logró concluir que el 52,22% (773 empresas) de las empresas testeadas se vieron afectadas por las siguientes variables: adquisición de maquinaria y equipo en los años 2017 y 2018, y que fueran pertenecientes a las actividades económicas dentro del sector alimenticio, textil, químicos, plásticos y de metales. Por lo cual se concluye que durante este periodo de tiempo lo que predomino fueron los valores monetarios invertidos por la empresa y la financiación de recursos públicos por encima de factores como la adquisición de maquinaria para mejoramiento de procesos o el nivel educativo de los empleados de la organización.

Afirmando lo que indico Manrique et l., (2019) es importante que se le dé prioridad a la innovación dentro de la cadena logística entendiendo que su adecuada gestión ayuda a potencializar y comercializar el producto haciendo que sea competitivo dentro del mercado, por lo cual se espera que el porcentaje de empresas testeadas en un futuro sea mayor y no solo un poco más de la mitad, pues se sobreentiende que el resto de las empresas no están generando un valor agregado a su producto y no son competitivas dentro del mercado, haciendo que sean cada vez más obsoletas y no están en la vanguardia de las tendencias mundiales como lo indica la CEPAL (2019).

Es importante recalcar, que la innovación no se realiza necesariamente con alguna inversión en maquinaria o valor monetario sino con diversas técnicas que puedan lograr un mejoramiento en la cadena logística, ya sea el rediseño de una ruta de entrega (como lo indicado por Henríquez-Fuentes et al., 2018) o el mejoramiento de una parte del proceso que pueda crear un valor agregado al producto y que generé beneficios dentro de la empresa. Adicionalmente, el personal capacitado tiene un papel fundamental para que se apliquen los conocimientos adecuados en los sistemas operativos y logren los objetivos organizacionales de la compañía.

Algunas limitaciones dentro del desarrollo de este estudio, fue que no se contaba con los datos más recientes de la encuesta lo que delimito los resultados encontrados.

Finalmente, se sugiere que para futuras investigaciones se emplee otro tipo de técnicas cuantitativas para analizar los mismos fenómenos de una manera más profunda y detallada teniendo en cuenta las nuevas tendencias mundiales y como ha afectado la cadena logística internacional.

Referencias

García, L. A. M. (2016). Gestión Logística Integral: Las mejores prácticas en la cadena dabastecimiento (2a ed.). Ecoe Ediciones.

ANDI (2016) Noticias. https://www.andi.com.co/Home/Noticia/15881-el-2020-fue-el-ano-de-la-aceleracion-de. Accessed 11 Nov. 2022.

OCDE (2005). The Measurement of Scientific and Technological Activities. Proposed guidelines for Collecting and Interpreting Technological innovation data. OSLO MANUAL. European Commission. Eurostat.

Estrada, Liriam. «Logística y tecnología de información como elementos de competitividad». Marketing Visionario, vol. 5, n.o 1, mayo de 2016, pp. 58-71. ojs.urbe.edu, http://ojs.urbe.edu/index.php/market/article/view/2390.

What is machine learning? (s. f.). Recuperado 15 de noviembre de 2022, de https://www.ibm.com/cloud/learn/machine-learning

Camacho Camacho, H., Gómez Espinosa, K., & Monroy, C. (2012). Importancia de la cadena de suministros en las organizaciones. Megaprojects: Building Infrastructure by fostering engineering collaboration, efficient and effective integration and innovative planning. Panama City: Tenth LACCEI Latin American and Caribbean Conference (LACCEI'2012).

Jasp (JASP 0.16.3). (2022). https://jasp-stats.org/download/

Departamento Nacional de Planeación. (2020). Observatorio Nacional de Logística, Transporte, Minas y Energía. Obtenido de Encuesta Nacional Logística: https://planeacionnacional.sharepoint.com/sites/PlataformaDIES2/Shared%20Documents/Encuesta%20Nacional%20Log%C3%ADstica/ENL%202020/ENL_2020_Documento_Resultados_10-08-2021(3).pdf

Reyes Bareño, P. A., & Valero Ortega, S. J. (2019). DHL: Innovación globalizada. Universidad del Rosario. https://repository.urosario.edu.co/bitstream/handle/10336/19805/ValeroOrtega-SilviaJuliana-1-2019.pdf?sequence=1&isAllowed=y

Martínez, L., & Kadi, O. E. (2019). Logística integral y calidad total, filosofía de gestión organizacional orientadas al cliente. Revista Arbitrada Interdisciplinaria Koinonía, 4(7 (Enero-Junio)), 202-232. https://dialnet.unirioja.es/servlet/articulo?codigo=7062704

Pérez, Gabriel, y Ricardo J. Sánchez. «Logística para la producción, la distribución y el comercio». CEPAL Transporte, vol. 1, 2019, pp. 9-10, https://repositorio.cepal.org/bitstream/handle/11362/44897/1/S1900719_es.pdf.

De la Hoz Granadillo, E., Martínez Sierra, D., Orozco Acosta, E., De la Hoz Reyes, R., Herrera Vega, J. C., Hernández Palma, H., Villanueva Cantillo, J., Ruiz Ohlsen, A., Ortiz Ospino, L. E., Mejía Acuña, F. J., Cervera Cárdenas, J. E., & Osío Ospino, R. C. (2017). Estudios de competitividad y análisis empresarial en la región Caribe. Ediciones Universidad Simón Bolívar. https://bonga.unisimon.edu.co/handle/20.500.12442/2953

Cano Olivos, P., Orue Carrasco, F., Martínez Flores, J. L., Mayett Moreno, Y., & López Nava, G. (2015). Modelo de gestión logística para pequeñas y medianas empresas en México. Contaduría y administración, 60(1), 181-203.

Robayo Acuña, P. V. (2016). La innovación como proceso y su gestión en la organización: Una aplicación para el sector gráfico colombiano. Suma de Negocios, 7(16), 125-140. https://doi.org/10.1016/j.sumneg.2016.02.007

Henríquez-Fuentes, G. R., Cardona, D. A., Rada-Llanos, J. A., Robles, N. R., N. R. (2018). Medición de Tiempos en un Sistema de Distribución bajo un Estudio de Métodos y Tiempos. Información tecnológica, 29(6), 277-286. https://doi.org/10.4067/S0718-07642018000600277

Jordan, V., Galperín, H., & Peres, W. (2013). Banda ancha en América Latina: Más allá de la conectividad. CEPAL. https://repositorio.cepal.org/bitstream/handle/11362/35426/S2013070_es.pdf

COVID-19: ¿Una oportunidad para la transformación digital de las pymes? (2020, abril 29). Puntos sobre la i. https://blogs.iadb.org/innovacion/es/covid-19-oportunidad-transformacion-digital-pymes/

Secretaria General Iberoamericana. (2021). *La Transformación Digital*. 31-37. https://www.andi.com.co/Uploads/INFTD.pdf

Muñoz, V., Hormechea, C. (2018-2020). ¿Cómo vamos avanzando en la transformación digital? Recuperado 15 de noviembre de 2022, de https://dapre.presidencia.gov.co/TD/Como-vamos-avanzando-en-la-TD-070421.pdf

Jadhav, A., Orr, S., & Malik, M. (2019). The role of supply chain orientation in achieving supply chain sustainability. *International Journal of Production Economics*, 112-125.

Forslund, H., Björklund, M., & Svensson Ülgen, V. (2021). Challenges in extending sustainability across a transport supply chain. Emerald Group Holdings Ltd., 1-16.

Awawdeh, H., Abulaila, H., Alshanty, A., & Alzoubi, A. (2021). Digital entrepreneurship and its impact on digital supply chains: The mediating role of business intelligence applications. International Journal of Data and Network Science, 233-242.

Manrique Nugent, M. A., Teves Quispe, J., Taco Llave, A. M., & Flores Morales, J. (2019). Gestión de cadena de suministro: una mirada desde la perspectiva teórica. Revista Venezolana de Gerencia, 1136-1146.

Budet Jofra, X., & Pérez Gómez, A. (2018). Innovaciones tecnológicas en la cadena de suministro aplicadas al eCommerce. Oikonomics Revista de los Estudios de Economía y Empresa, 41-57.

IBM Institute for Business Value. (20 de Septiembre de 2022). CSCO Study: Achieving data-led innovation. Obtenido de IBM Institute for Business Value: https://www.ibm.com/thought-leadership/institute-business-value/en-us/c-suite-study/csco

Python Software Foundation. (2022). El tutorial de Python. Obtenido de Python: https://docs.python.org/es/3/tutorial/

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). Tree-Based Methods. En An Introduction to Statistical Learning (págs. 311-314). Nueva York: Springer New York, NY.

DANE. (2020). Microdatos. Obtenido de DANE (Departamento Administrativo Nacional de Estadística): https://microdatos.dane.gov.co/catalog/651/study-description

Anexos

Anexo 1

CIIU	Actividad Económica
1011	Procesamiento y conservación de carne y productos cárnicos
1051	Elaboración de productos de molinería
1081	Elaboración de productos de panadería
1082	Elaboración de otros productos alimenticios (Cacao, chocolate y productos de confitería)
1083	Elaboración de macarrones, fideos, alcuzcuz y productos farináceos similares
1089	Elaboración de otros productos alimenticios n.c.p.
1090	Elaboración de alimentos preparados para animales
1101	Destilación, rectificación y mezcla de bebidas alcohólicas
1102	Elaboración de bebidas fermentadas no destiladas
1104	Elaboración de bebidas no alcohólicas, producción de aguas minerales y de otras aguas embotelladas
1392	Fabricación de otros productos textiles (Confección de articulos con materiales textiles, excepto prendas de vestir)
1393	Fabricación de tapetes y alfombras para pisos
1399	Fabricación de otros productos textiles (Fabricación de otros artículos textiles n.c.p.)
1410	Confección de prendas de vestir, excepto prendas de piel
1512	Fabricación de artículos de viaje, bolsos de mano y artículos similares elaborados en cuero, y fabricación de artículos de talabartería y guarnicionería
1513	Fabricación de artículos de viaje, bolsos de mano y artículos similares; artículos de talabartería y guarnicionería elaborados en otros materiales
1521	Fabricación de calzado de cuero y piel, con cualquier tipo de suela
1523	Fabricación de partes del calzado
1640	Fabricación de recipientes de madera
1709	Fabricación de papel, cartón y productos de papel y cartón
1910	Fabricación de productos de hornos de coque
2011	Fabricación de sustancias y productos químicos básicos
2023	Fabricación de jabones y detergentes, preparados para limpiar y pulir; perfumes y preparados de tocador

Anexo 2

CIIU	Actividad Económica
1011	Procesamiento y conservación de carne y productos cárnicos
1040	Elaboración de productos lácteos
1051	Elaboración de productos de molinería
1071	Elaboración y refinación de azúcar
1081	Elaboración de productos de panadería
1090	Elaboración de alimentos preparados para animales
1101	Destilación, rectificación y mezcla de bebidas alcohólicas
1104	Elaboración de bebidas no alcohólicas, producción de aguas minerales y de otras aguas embotelladas
1311	Preparación e hilatura de fibras textiles
1313	Acabado de productos textiles
1410	Confección de prendas de vestir, excepto prendas de piel
1521	Fabricación de calzado de cuero y piel, con cualquier tipo de suela
1522	Fabricación de otros tipos de calzado, excepto calzado de cuero y piel
1523	Fabricación de partes del calzado
1610	Aserrado, acepillado e impregnación de la madera
1620	Fabricación de hojas de madera para enchapado; fabricación de tableros contrachapados, tableros laminados, tableros de partículas y otros tableros y paneles
1701	Fabricación de pulpas (pastas) celulósicas; papel y cartón
1921	Fabricación de productos de la refinación del petróleo
1922	Actividad de mezcla de combustibles
2021	Fabricación de plaguicidas y otros productos químicos de uso agropecuario
2023	Fabricación de jabones y detergentes, preparados para limpiar y pulir; perfumes y preparados de tocador
2212	Reencauche de llantas usadas
2221	Fabricación de formas básicas de plástico
2310	Fabricación de vidrio y productos de vidrio
2393	Fabricación de otros productos de cerámica y porcelana
2394	Fabricación de cemento, cal y yeso
2395	Fabricación de artículos de hormigón, cemento y yeso
2410	Industrias básicas de hierro y de acero
2429	Industrias básicas de otros metales no ferrosos
2431	Fundición de hierro y de acero
2511	Fabricación de productos metálicos para uso estructural
2513	Fabricación de generadores de vapor, excepto calderas de agua caliente para calefacción central
2593	Fabricación de artículos de cuchillería, herramientas de mano y artículos de ferretería
2599	Fabricación de otros productos elaborados de metal n.c.p.

2610	Fabricación de componentes y tableros electrónicos
2711	Fabricación de motores, generadores y transformadores eléctricos

Anexo 3

CIIU	Actividad Económica
1012	Procesamiento y conservación de pescados, crustáceos y moluscos.
1052	Elaboración de almidones y productos derivados del almidón.
1061	Elaboración de productos de café (Trilla de Café)
1062	Elaboración de productos de café (Descafeinado, tostón y molienda del café)
1082	Elaboración de otros productos alimenticios (Cacao, chocolate y productos de confitería)
1089	Elaboración de otros productos alimenticios n.c.p.
1312	Fabricación de productos textiles (tejeduría de productos textiles)
1391	Fabricación de otros productos textiles (tejidos de punto y ganchillo)
1392	Fabricación de otros productos textiles (Confección de artículos con materiales textiles, excepto prendas de vestir)
1394	Fabricación de otros productos textiles (Fabricación de cuerdas, cordeles, cables, bramantes y redes)
1399	Fabricación de otros productos textiles (Fabricación de otros artículos textiles n.c.p.)
1430	Confección de prendas de vestir (Fabricación de artículos de punto y ganchillo)
1630	Transformación de la madera y fabricación de productos de madera y de corcho, excepto muebles; fabricación de artículos de cestería y espartería
1709	Fabricación de papel, cartón y productos de papel y cartón
2011	Fabricación de sustancias y productos químicos básicos
2012	Fabricación de abonos y compuestos inorgánicos nitrogenados
2013	Fabricación de plásticos en formas primarias
2022	Fabricación de pinturas, barnices y revestimientos similares, tintas para impresión y masillas
2029	Fabricación de otros productos químicos n.c.p.
2100	Fabricación de productos farmacéuticos, sustancias químicas medicinales y productos botánicos de uso farmacéutico
2219	Fabricación de formas básicas de caucho y otros productos de caucho n.c.p.
2229	Fabricación de artículos de plástico n.c.p.
2392	Fabricación de materiales de arcilla para la construcción
2399	Fabricación de otros productos minerales no metálicos n.c.p.
2592	Fabricación de otros productos elaborados de metal (Tratamiento y revestimiento de metales; mecanizado)
2651	Fabricación de equipo de medición, prueba, navegación y control
2652	Fabricación de relojes
2712	Fabricación de aparatos de distribución y control de la energía eléctrica
2720	Fabricación de pilas, baterías y acumuladores eléctricos
2740	Fabricación de equipos eléctricos de iluminación

2819	Fabricación de otros tipos de maquinaria y equipo de uso general n.c.p.
2910	Fabricación de vehículos automotores y sus motores
2920	Fabricación de carrocerías para vehículos automotores; fabricación de remolques y semirremolques
3240	Fabricación de vehículos militares de combate