Modelling Experiences of the COVID-19 Pandemic in Ibero-America

Modelling Experiences of the COVID-19 Pandemic in Ibero-America

Edited by

Carlos N. Bouza-Herrera

Cambridge
Scholars
Publishing

Modelling Experiences of the COVID-19 Pandemic in Ibero-America

Edited by Carlos N. Bouza-Herrera

This book first published 2023

Cambridge Scholars Publishing

Lady Stephenson Library, Newcastle upon Tyne, NE6 2PA, UK

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

ISBN (10): 1-5275-0719-X
ISBN (13): 978-1-5275-0719-7

TABLE OF CONTENTS

AUTHOR

Figure 1: Dr. Carlos N. Bouza Herrera

We are pleased to present the editor–prof. Dr. Carlos N. Bouza-Herrera (CB). He was born and lived in La Habana, Cuba, where he studied at public schools. At the age of 17 he began working at a Cuban Telephone Company and continued his secondary school studies. CB got involved in the tide of political changes that occurred in his country and stopped studying for a while. In 1966 he matriculated in the School of Mathematics of Universidad de La Habana and selected the specialization in Mathematical Statistics. He shared working, studying and politics.

The new Cuban society developed a series of research institutions in medicine and agriculture. The country was urged to have statisticians duly trained. The different university careers now included the teaching of statistics in their curriculums and there was a lack of such teachers. Advanced students of Mathematical Statistics were encouraged to teach and CB got involved, and included these activities in his daily life. Once graduated, he was ubicated in the corresponding department of the School of Mathematics.

Research institutes of the country needed statistical aids, and projects with them were promoted systematically. CB specialized in modeling survey sampling for biologists, engineers and sociologists, mainly. These

applications produced several reports as early as 1973. A PNUD-UNESCO project supported the preparation of statistics and computer sciences professors. Prof. Dr. Vitomir Erdeljan was in charge, for 2 years, of working with statisticians. CB was one of the trainees and under his advisory obtained his MSC (1974) in Cuba and PhD (1978) in the ancient Yugoslavia.

CB has directed 69 projects, of which, 2 are still running. He had visited more than 50 institutions around the world, and particularly he made various visits to the Indian Statistical Institute of Kolkata supported by the Third World Academy of Sciences, where he started a long-term cooperation with generations of Indian statisticians. To our knowledge, he has published more than 270 papers, edited or authored more than 30 books, and participated in more than 150 congresses. CB is member of 15 professional societies and networks and acts as a referee or member of the boards of 51 journals all around the world. CB has been awarded with more than 50 awards with diplomas and medals in his country.

AD-Scientific Index - World Scientist Rankings – 2023 reports the scientific impacts of him as follows:

H Index: Total	17, Last 5 years 13	total: 0,765
i10 INDEX Total	Last 5 year	Last 5 years 17, total 0,708
CITATIONS: Total 910	Last 5 years 555	total 0,610

ADSI ranked CB as 33rd in his university, and 113th among the scientists in Cuba.

Our collaboration with CB has a long history. He participated as a teacher in our MSC program and as an advisor of MSC and PhD theses and other research. Collaborating with other institutions, he has been a motivation for developing joint research within networks. Red iberoamericana de Estudios Cuantitativos Aplicados is one of them. This network connects Ibero-America's researchers and French's institutions collaborators and, and they produce yearly a book with results of the cooperation. Note that, though he is pensioned, CB continues actively working both in teaching and investigating, having the leadership of programs, projects and several activities where his experience is needed. This book has been provoked by the work of collaborators of CB on the study of Covid-19.

CB is happily married with Sira Allende and they have a daughter, Gemayqzel. They are mathematicians too. As a person, his sense of humor is proverbial, in good and bad times. That stimulates his colleagues with looking for a solution, when the expected results are not satisfactory. Distressing is innate and in the joint work with his colleagues, friends and relatives, his advises come not as reprimands but as a product of some philosophical reflection. He is prompt to aid everyone, not only in the classrooms or at the research laboratories but also in current life. His experience in living provides to people ideas for solving seemingly too cumbersome problems that new generations have to deal with. We and his network congratulate him and express our wishes of a long life.

Agustin Santiago-Moreno and Jose Maclovio Sautto-Vallejo.

PROLOGUE

The COVID-19 pandemic has posed a set of unprecedented challenges to humanity. The management of healthcare resources should be based on quantitative studies as it has generated a lot of epidemic data. Quantitative modelling allows generalizing which can guide the decision makers. The impact of COVID-19 in healthcare resources is particularly affecting lower and middle-income countries. Latin American countries are disadvantaged due to the lack of medical personal, beds in hospitals and medical supplies. Nonmedical specialists are contributing to the development of quantitative models, which are aiding the study of the COVID-19 pandemic dynamics. Some countries have structured public policies using some mathematical models.

Nowadays, quantitative models, mathematical and simulation mainly, are used for planning and/or evaluating health issues. Recommendations based on quantitative studies improves strategic, tactical, and operational aspects of preparedness planning in the management of the pandemic.

The recommendations derived from the research that generated the chapters of this oeuvre are a significative sample of studies developed in Latin America on the COVID-19 pandemic during 2020.

Chapter I is concerned with the development of an ordinary differential equation model and formulates the expression of Basic Reproductive Number (R_0) for the subpopulation of undetected infected individuals. The authors modelled risk perception through proposed functions and simulated governmental actions and individual attitudes, with a contagion variable rate over time. Real data were used for predicting the behaviour of the pandemic.

Chapter II proposes, discusses and illustrates Agent-Based Modelling to handle complexity and its consequent uncertainty in decision-making processes, based on the law of the dynamic equilibrium of markets and the unlimited rationality of both social and economic agents. They explored and estimated how the establishment of rules for the social behaviour of a community affects the dynamics of an epidemic. An application was developed, in NetLogo, that allows the construction of a model of a city

from its basic demographic information: size, population, inhabitants per house, number of closed public areas and amount of free area.

Chapter III is devoted to describing the behaviour of COVID-19 in Mexico from the first confirmed cases on February 28, 2020, until May 2021. This characterization of COVID-19 in Mexico was made through positive cases, cases with comorbidities, number of deaths, mortality rate, incidence rate and fatality rate; represented by state, sex and age group. The analysis of the behaviour of the pandemic in Mexico used the correlation coefficients among three relevant variables against two indices-summaries, for the data of the states. The relevant COVID-19 variables considered were incidence, mortality rate, and fatality rate. The relations analysed were the Marginalization Index and the Human Development Index studied under the conditions of the pandemic, considering the socioeconomic conditions of the states of Mexico.

Chapter IV has considered a supervised multilayer neural network – Artificial Neural Network – for fitting the data of estimated cases of COVID-19 in the State of Puebla. However, this model was inefficient in several cases. The authors developed alternatively a "vectorial" neural network with an unvarying step function activity. A positive defined matrix allowed training the variable learning rate model, using the existing data on the disease, from week 10 of 2020 to week 35 of 2021.

Chapter V has proposed models for post-pandemic studies and the sequels of the treatments. Theoretical models for epidemic graphs were characterized for aiding the epidemiologists in evaluating the effect of treatments (sequels, correlations with demographic variables, etc.). The modelling in the context of COVID-19 was developed.

Chapter VI has considered problems on sampling in the presence of missing observations. The authors established regularities of the expectations under a particular non-response mechanism, considering some issues present in real life statistical research in COVID-19. Imputation procedures and superpopulation modelling have provided a theoretical frame for dealing with missing observations within the framework of sampling, looking for identifying contacts with infected people.

Chapter VII poses a discussion on the effects of the COVID-19 pandemic in increasing the visible inequality in Latin American countries. The authors have developed a study in an especially vulnerable sector of indigenous Mexican students.

Chapter VIII presents a study on hospital-staying in surviving individuals from COVID-19 of Cuban recovered patients living in Havana. The article proposes future lines of research.

In Chapter IX, a prospective pilot study is developed for establishing the behaviour of recovered COVID-19 patients with lung damage treated with a new medicament.

This book provides a good example of the research developed by Latin American scientists for dealing with the challenges posed by the COVID-19 pandemic.

CHAPTER I

MANAGEMENT OF COVID-19 IN CUBA WITH RISK PERCEPTION: A SAIR-TYPE MODEL CONSIDERING THE ASYMPTOMATIC POPULATION

DANIEL MENCIÓ PADRÓN[1], GABRIELA BAYOLO SOLER[2] AND AYMÉE MARRERO SEVERO[2]

Summary

Based on epidemic population models defined by ordinary SAIR-type differential equations, this work presents a variant that distinguishes between the populations of uncontrolled symptomatic and asymptomatic infected people, moving freely in society, to represent the transmission dynamics of COVID-19 and the subpopulation of individuals in quarantine and hospitalized, managed by the institutions of the Health System.

It presents the generated diagram for the population model defined by ordinary differential equations. Also, it formulates the expression of Basic Reproductive Number (R_0) for the subpopulation of infected individuals that are not detected, whom are, consequently, the main ones guilty of the transmission of the epidemic.

It shows the obtained results by fitting essential parameters of the models by considering data of Cuba in the first 51 days of 2021. The formulation of risk perception through functions that simulate government actions and

[1] School of Economics, University of Havana, Havana, Cuba.
[2] Department of Applied Mathematics, University of Havana, Havana, Cuba.

individual attitudes, with a variable contagion rate over time, allowed us to simulate different scenarios by considering epidemic waves and present useful predictions for efficient actions of the Public Health System and other authorities in order to control the pandemic.

Keywords: epidemic model, SAIR-type, asymptomatic population, infection rate, risk perception.

1. Introduction

In Cuba, the first cases of SarsCov2 infection were reported on March 11th, 2020. Three foreigners arrived in the country from Italy. Before that, the Cuban Ministry of Health and the Cuban scientific community had accelerated studies and strategies addressed to declare COVID-19 a pandemic in the country.

Works presented by researchers from countries around the world provided highly useful information for mathematical modelling. Taking into account the so-called social distancing or isolation as an efficient way to control it, reducing person-to-person contact with the aim of controlling the spread, avoiding high numbers of infected and sick people and, therefore, decreasing the impact and extent of the disease (Gutiérrez and Varona 2020), (Lin and et al 2020), (López 2006), after one year, there is no specific curative treatment to prevent the collapse of health services, either due to lack of resources, both human and material, or due to the complexity of access to vaccines that generate proven immunity; therefore, studies of this nature maintain their value and importance.

The Cuban scientific community has provided a considerable number of mathematical models to describe the transmission dynamics of this disease, which have contributed to design State and Health System control protocols to deal with it.

The papers Chen (2020), Gutiérrez and Varona (2020), Lin and et al. (2020) and Wu, Leung and Leung (2020) motivated the authors of this work to make the initial proposals Marrero, Menció and Bayolo (2020) and Menció, Bayolo and Marrero (2020) variants, called SEAIR and SAIRV, of the classic population type SEIR. The SEAIR model adds the subpopulation of individuals in the incubation or latency period, distinguishing between the symptomatic and asymptomatic infected, considering this latter as a new subpopulation; the SAIRV model considers a variant with exposure to the

virus and the differentiation of asymptomatic and infected people hospitalized, quarantined and moving freely in society.
The results obtained and the intention to refine details that explain and characterize the endemic waves or new outbreaks that occurred in Cuba motivated the proposal of a new SAIR-type variant that, as in the previous models, considers variable transmission rates based on parameters that represent the strength of the control action and the risk perception, keeping, for some of the parameters, the definitions of previous works (Marrero, Menció and Bayolo 2020), (Menció, Bayolo and Marrero 2020). An adjustment of the main parameters that characterize the transmission dynamics and the estimates has been carried out using MATLAB version R2018a, which allowed validation and comparison of results.

2. Mathematically modelling COVID-19 with risk perception

This work shall intend to analyse an infectious disease, overall, in its ability to invade a population.

To describe this effect mathematically, systems of ordinary differential equations are used, among other tools.

The conditions of existence and uniqueness of the solution of a system of ordinary differential equations are based on the classical theorem (Elsgoltz 1969). However, its great applicability to disease transmission processes has supported the formulation of an epidemiological variant of this theorem, valid in a biologically feasible region, which guarantees the non-negativity of the solutions.

Theorem. Let $F: \mathbb{R}_+^n \to \mathbb{R}_+^n$ be a locally continuous Lipchitz function and assume that $F_j(x) \geq 0, \forall x \in \mathbb{R}^n \; \forall j = 1,2, \dots, n$. Then, $\forall x_0 \in \mathbb{R}^n$, there is a unique solution of $\dot{x} = F(x), x(0) = x_0, \;\; x_0 \in \mathbb{R}_+^n$, x_0 belonging to the interval $[a, b]$ with $b \in [0, \infty)$. Demonstration in (López 2006).

These theoretical mathematical results guarantee that the proposed model makes epidemiological sense, within an invariant region in which all the solutions of the system remain non-negative and bounded, $\forall \; t > \; 0$.

Particularly, in epidemiological models the existence of the infection-free point (in which there is an absence of disease in the population) and a threshold parameter called Basic Reproductive Number (R_0) are both taken

into consideration. The (R_0) has great importance in the analysis of transmission. If $R_0 < 1$, infection-free point is asymptotically locally stable, therefore the disease does not invade the population. This indicator could be understood so that, on average, an infected individual produces less than one new infected individual in the course of their infectious period, so the disease does not grow. If $R_0 > 1$, infection-free point is unstable and disease will probably invade the population, since every infected individual produces on average more than one new case, and so disease may become an epidemic (infected population number increases); when $R_0 = 1$, it denotes disease endemism or, at least, permanence in time.

The novelty of this disease is assumed; its short time of development around the world; and the lack of certainty in its medical-clinical and epidemiological behaviour. Therefore, each modelling proposal must take certain hypotheses, essentially related to uncertainty about, for example, period of infectiveness characteristics, antibody count, differentiation of contagion rates between asymptomatic and symptomatic, period of immunity after recovering from the disease, etc.

Based on various international articles (Chen 2020), (Gutiérrez and Varona 2020), (Tang and et al 2020), (Wu, Leung and Leung 2020) and our previous proposals (Marrero, Menció and Bayolo 2020), (Menció, Bayolo and Marrero 2020), we present in this work a new SAIR-type variant with differentiation of free and controlled populations, assuming temporary immunity after suffering the disease and considering variable transmission rates based on parameters that represent the strength of the control action and the perception of risk.

2.1 Model variables

$S\ (t)$: Susceptible population (healthy people in the population at time t)

$A_L\ (t)$: Asymptomatic free population (infected people, not tested at time t, who do not develop symptoms and remain in the population until they are detected or pass the period of infestation)

$I_L\ (t)$: Free infected population (infected people, not tested at time t, who develop symptoms and remain in the population until they are detected, recover or die).

$Q\ (t)$: In quarantine population (people, infected or not, who are confined at time t, either because they are contacts of positive cases or because they are suspected of being infected, until they are tested)

$I_h\ (t)$: Hospitalized infected population (infected people, tested at time t, who remain in this subpopulation until they recover or die)

$R\ (t)$: Recovered population (all people who, at time t, after the infestation period, have recovered from the disease, whether or not they have been hospitalized).

Health authorities do not control subpopulation sS, A_L, I_L, so it may be considered that they are transmitting the virus in society.

Subpopulations Q and I_h are under the control of health and community entities, either in hospitals, isolation centers or under observation at their homes.

It is important to emphasize that, in our conception of recovered people, not only those controlled by health entities, but also those infected (asymptomatic and symptomatic), who freely develop the disease, are taken into account. Another characteristic of the model is that, taking into account the Cuban strategy of management and control of the disease, those recovered with negative tests are returned to confinement for a defined period.

2.2 The model

$$\frac{dS}{dt} = -\frac{S}{N}\left[\beta(t)(I_L + \theta A_L + cP_f) + \beta_c Q + \beta_h I_h\right] - \sigma S + \left(1 - d_Q\right)\tau_Q Q + (1-c)P_f$$

$$\frac{dA_L}{dt} = \frac{S}{N}\beta(t)(1-q)(I_L + \theta A_L + cP_f) - (d_{AL} + \tau_{AL})A_L$$

$$\frac{dI_L}{dt} = \frac{S}{N}\beta(t)q(I_L + \theta A_L + cP_f) - (d_{IL} + m + \tau_{IL})I_L$$

$$\frac{dQ}{dt} = \frac{S}{N}(\beta_c Q + \beta_h I_h) + \sigma S - \tau_Q Q + \gamma_h I_h - \tau_R \gamma_h I_h$$

$$\frac{dI_h}{dt} = d_Q \tau_Q Q + d_{AL} A_L + d_{IL} I_L - (m + \gamma_h) I_h$$

$$\frac{dR}{dt} = \tau_{AL} A_L + \tau_{IL} I_L + \tau_R \gamma_h I_h$$

2.3 Diagram of the model

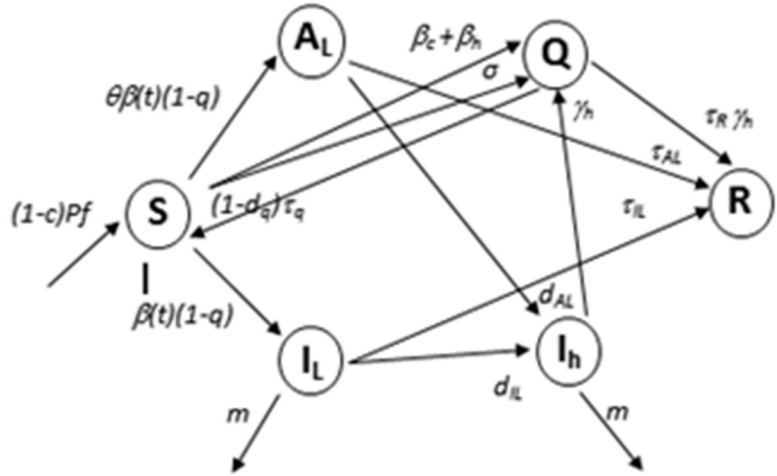


Fig. 2-1. Diagram of the model

2.4 About the Basic Reproductive Number R_0

To calculate the indicator R_0 of the model proposed in this work, the next generation matrix was used, a tool widely reported in the literature for its simplicity and viability (Marrero, Menció and Bayolo 2020). It was obtained that, at the infection-free point, all the variables take a value of zero except S_0, the susceptible population at the initial instant, which coincides with the total population.

Having in mind the characteristics of this disease and its behaviour generate epidemic waves or regrowth, an expression to R_0 variable over time was obtained. This aspect is conditioned essentially by its dependence (among others) on the transmission rate, which is also variable. In the procedure developed to the calculation of R_0, the parameters that determine transmission in subpopulations of free infected people (asymptomatic and symptomatic) and in subpopulations of controlled infected people (in quarantine and hospitalized) were discriminated. This determines that the greatest source of contagion lies in the free infected subpopulation. The following expression for R_0 parameter characterizes these subpopulations and strongly influences all transmission:

$$R_{0L}(t) = \beta(t)\left[\frac{(1-q)\theta}{(d_{AL}+\tau_{AL})} + \frac{q}{(d_{IL}+m+\tau_{IL})}\right]\frac{S_0}{N}$$

2.5 Model Parameters

P_f: Floating population entering the country. It is generated as a random number, bounded in a certain interval, taking into account the analysis of the available data.

σ: Proportion of suspects and people under observation. In the protocol for coping with COVID-19 in Cuba, the so-called "epidemiological spider" (also known as contact tracing) is established to isolate and control, during a period of time established by the characteristics of the disease, all the possible contacts of the detected infected person.

θ: Transmission constant for asymptomatic patients.

c: Proportion of infected people.

γ_h: Recovery rate in hospitalized patients who suffered from the disease

d_i: Positive detection rate for each subpopulation, for $i = Q, A_L, I_L$. It represents the proportion of individuals detected positive with confirmatory tests such as PCR, antigen test, etc., which motivates their confinement or hospitalization.

τ_Q: Infectiousness rate in inmates

τ_{AL}: Infectiousness rate in non-controlled asymptomatic people

τ_{IL} : Infectiousness rate in non-controlled infected people

τ_R : Infectiousness rate in recovered patient

With $\tau_i = \frac{1}{p_i}$ for $i = Q, A_L, I_L, R$; where p_i represents the infectious periods according to subpopulation.

m: Death rate from disease

The transmission rate of diseases under study is commonly considered as a more invariant parameter over time. However, in this model, as in the previous proposals (Marrero, Menció and Bayolo 2020), (Menció, Bayolo

and Marrero 2020), this rate is considered as a variable function over time, a result that also appears in recent works on this subject.

$\beta(t)$: Infection or transmission rate in free infected populations (asymptomatic and symptomatic)

$$\beta(t) = b_0(1-\alpha(t))\left(1 - m\left(\frac{I_h + I_L}{N}\right)\right)^k$$

$\alpha(t) = 1 - e^{-\delta t}$ represents the strength of the actions in the face of risk perception.

β_c, β_h: Contagion or transmission rate in confined and hospitalized populations respectively.

$$\beta_c(t) = b_{c0} * (1-\alpha)(1-\rho * dp_Q Q),\ \beta_h(t) = b_{h0} * (1-\alpha)(1-\rho I_h),$$

where ρ defines the coefficient of transmission adjustment.

3. Estimation of the model parameters

In obtaining acceptable results, it is indispensable to have reliable values of the essential parameters that characterize the transmission dynamics. For this reason, the parameter estimation problem for models described by ordinary differential equations is addressed, formulated as an optimization problem associated with finding the vector of optimal parameters in the model in question, minimizing the norm of the relative residual error between the actual data with the estimates obtained at each instant of time. Therefore, following the function, it's been defined:

$$e_i = \left\| \frac{X_{dat}(t_i) - X_{EDO}(t_i)}{X_{dat}(t_i)} \right\|_2, i = 1, \dots, n$$

to minimize $W = \sum_{i=1}^{n} w_i e_i$, where w_i represents the weights, $X_{dat}(t_i)$ vector of data from the variables of the model at each instant of time and $X_{EDO}(t_i)$ the corresponding estimates.

In this work, a strategy of hybridizing the Simulated Annealing heuristic with Quasi-newton methods, implemented in MATLAB, has been used to find the vector of optimal parameters, in order to refine the starting point for

the classical optimization methods with functions like *fminsearch* and *fmincom*.

Best results were obtained using MATLAB function *fminsearch*, considering the three variables of the model for which data were available and adjusting the model for the first 51 days of the epidemic in 2021. During adjustment, the function $\alpha(t)$ was defined as follows:

$$\alpha(t) = \begin{cases} 1 - e^{-\delta_1 t}, & 0 \le t \le 37 \\ \dfrac{1 - e^{-a_1(t-37)}}{a_2}, & 37 < t \le 41 \\ 1 - e^{-\delta_2 t}, & 41 \le t \le 48 \\ \dfrac{1 - e^{-a_1(t-41)}}{a_2}, & 48 < t \le 50 \\ 1 - e^{-\delta_1 t}, & t > 50 \end{cases}$$

From a numerical point of view, it is difficult to obtain valid results when the number of parameters is too large. Overall, comparing the number of variables in the model and the available data. Therefore, a certain number of parameters to be estimated is chosen and others are taken from the literature, as the opinion of specialists or according to reliable publications. Table 3.1 shows the parameters that were estimated in this investigation, the rest was taken from (Marrero, Menció and Bayolo 2020) and (Menció, Bayolo and Marrero 2020). The average value of the floating population was estimated to be around 455 people. Moreover, although the theoretical model discriminates between the detection rates according to the population, in the experiments d parameter was considered the same one for all the subpopulations.

Parameter	Value
d	0.0234451293424371
σ	0.00168771213544432
τ_Q	1.10742371672635
τ_{AL}	0.0248074364657216
τ_{IL}	0.0524177536853999

τ_r	0.595379115159391
b_0	0.0203526228007360
b_{h0}	0.000515051376863783
b_{c0}	3.81591585158797
δ_1	0.138436211668359
δ_2	0.0682671745032089
a_1	-3.94988438426982e-6
a_2	2.12238077138776

Table 3-1. Estimated parameters.

4. Analysis of Results and Preliminary Conclusions

In previous works (Marrero, Menció and Bayolo 2020), (Menció, Bayolo and Marrero 2020), data from approximately the first three months of the pandemic in Cuba were used, a period in which the pandemic was successfully controlled with predictions for 150 days, when the beginning of a second regrowth began to manifest.

These results conditioned a new epidemic wave in the year 2021, among other factors, such as the reopening of international borders in Cuba in November 2020 and the festive activities at the end of the year. In this proposal, we handled the data of the first 51 days of the year 2021, considering $\alpha(t)$ as defined before, with the estimated values of δ parameter, which characterizes the actions based on risk perception and on which the transmission rate $\beta\ (t)$ essentially depends. This allowed us to simulate a new outbreak with figures much higher than in all of 2020 and to present estimates considering data for the first 51 days of the year 2021.

The figures 4.1, 4.2 and 4.3 show the results for Hospitalized, Quarantined and Recovered variables of the model for the first 51 days of this year. Those are consequence of the Parameter Estimation Problem explained above.

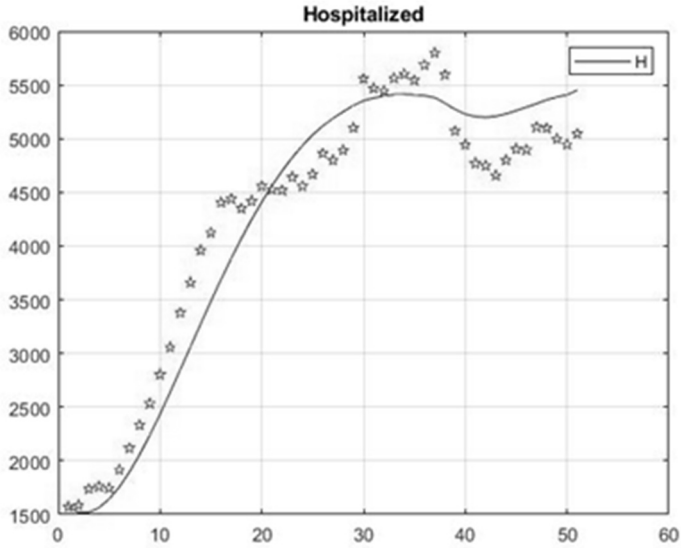


Fig. 4-1 Estimation of the hospitalized subpopulation at each instant of time. The red stars represent the data available in the first 51 days of the year 2021.

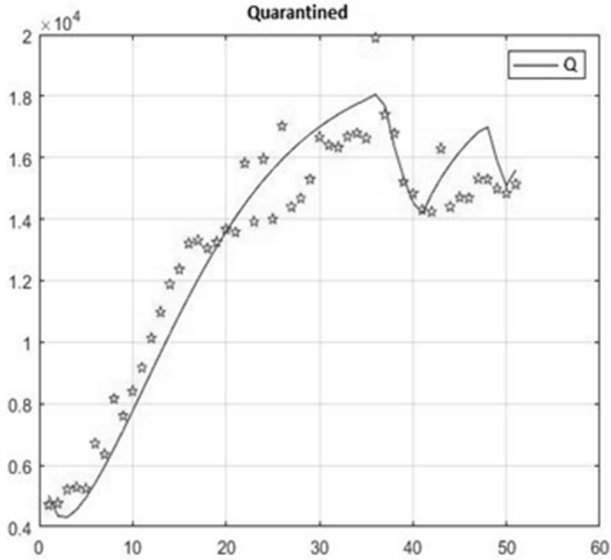


Fig. 4.2 Estimation of the subpopulation in quarantine at each instant of time The red stars represent the data available in the first 51 days of the year 2021.

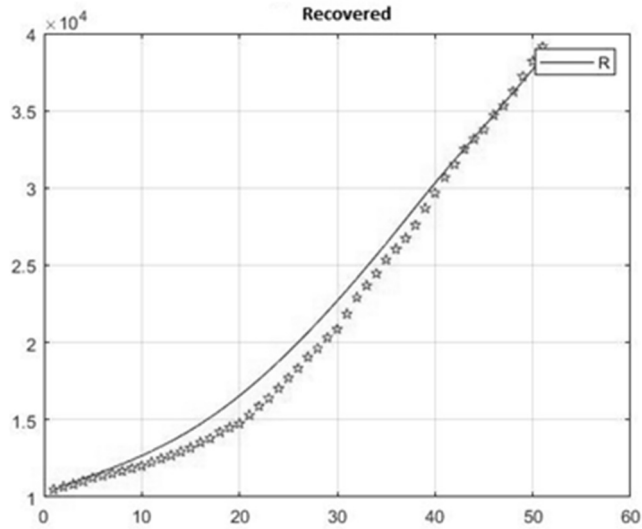


Fig. 4-3 Estimation of the recovered subpopulation at each instant of time. The red stars represent the data available in the first 51 days of the year 2021.

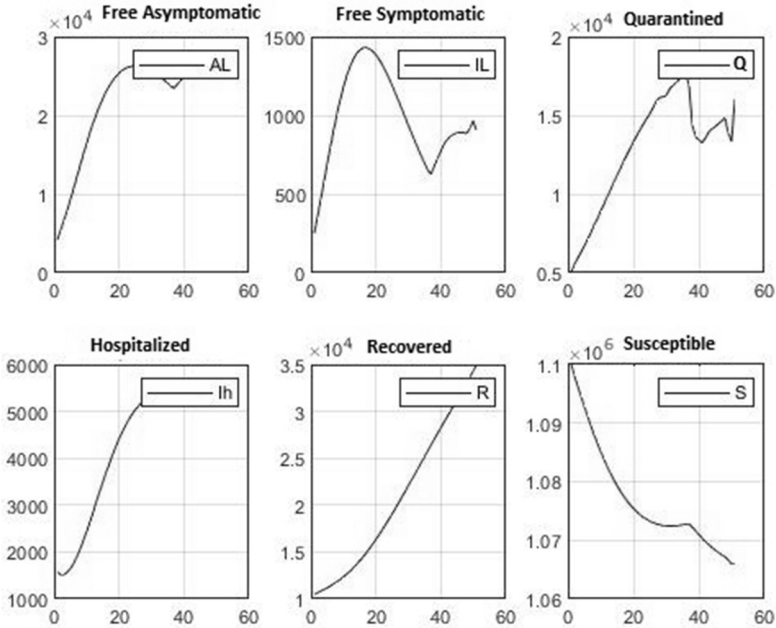


Fig. 4-4 Estimates for each of the model variables

An analysis of Figure 4.4 allows us to understand the influence of each variable on the behaviour of the others. Note that the A_L subpopulation reaches a maximum value close to 29 thousand. Therefore, it is the variable that reports the largest number of patients and yet it is the one with less information.

In order to predict the future behaviour of the hospitalized infected people variable, which is the most interesting for the health system, different possible scenarios were handled for this subpopulation in a period of 150 days after the analysed data. Different expressions were taken into account for the parameter $\alpha(t)$, which characterizes the actions based on risk perception and on which the transmission rate $\beta(t)$ depends, with the intention of simulating situations of greater and lower rigor in the measurements of control.

We worked with the following expressions of $\alpha(t)$, whose behaviours are represented in Figure 4.5.

$\alpha_1(t) = -t^{1/d}$, which simulates a scenario of greater control and perception of risk, therefore, the most favourable (red colour).

$\alpha_2(t) = -t^{1/d}$, which simulates a scenario of less control and perception of risk, therefore, the less favourable (green colour).

$\alpha_3(t) = asen(bt) + c$, which allows simulating a situation closer to the oscillations and regrowth that occur commonly (blue colour).

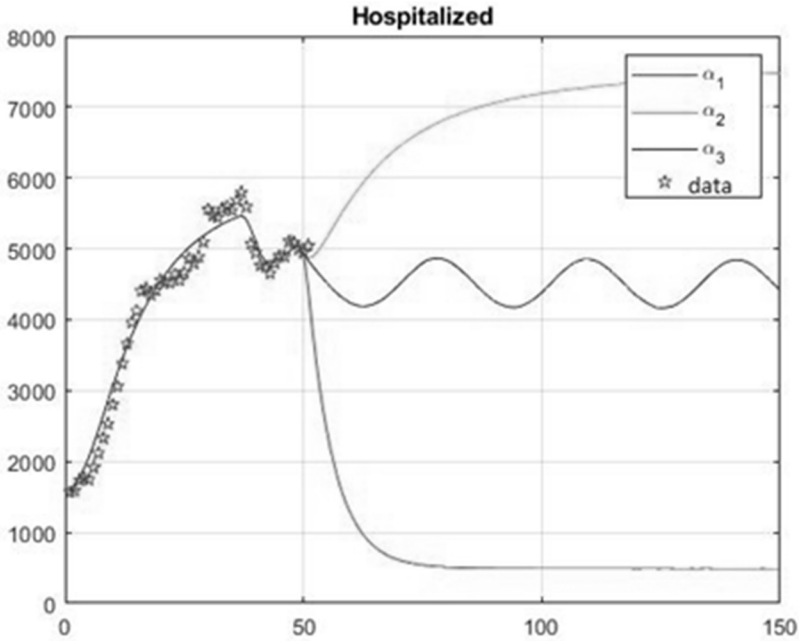


Fig. 4-5 Simulation of the hospitalized subpopulation for the next 150 days, based on the estimates corresponding to the first 51 days of the year 2021.

Using the Parameter Estimation Problem, the behaviour of model variables could be predicted, specially in the short and medium term. This proposal keeps in mind factors that pay to the perception of risk from the individual action as government measures, which are indispensable for the handling of the pandemic.

The different values of parameter $\alpha(t)$, explained previously, allow the simulation of various scenarios related to the perception of risk. In the figure 4.5, these situations are demonstrated for the Hospitalized Subpopulation, which is one of the most important for the Healthy Public System in our country.

The obtained results confirm that the control actions are the principal fact to manage the disease, although, according to the data of this outbreak, a number of sick people will certainly always remain.

5. Final remarks

In this work, a SAIR-type model is presented to simulate the transmission dynamics of COVID-19. Particularly, it subdivides the asymptomatic and infected populations that are non-controlled in society and differentiates between confined and hospitalized people from those who are controlled and attended by the instances of the Cuban Public Health System. During modelling, some of the precepts that characterize the protocol in coping with the disease in our country were represented.

The Parameter Estimation Problem was solved to find the optimal values of the main parameters that characterize the transmission dynamics. To achieve this, a classic optimization problem was formulated, with a minimum-square objective function weighted by the lack of data for the variables of free infected population, both symptomatic and asymptomatic. The satisfactory results obtained with respect to the data of the first 51 days of the year 2021 allowed the simulation of and infer the immediate future behaviour of the disease, in a period of 150 days for the variable hospitalized infected population.

A valid characteristic of the presented proposal is to consider the transmission rate variable and dependent, among others, on factors that simulate the strength of government actions, of health entities and the risk perception of the population. This explains the sensitivity of the results to the variations of these factors. In the case of the hospitalized infected subpopulation, strategies have been used to allow simulating different behaviours, depending precisely on the strength of the control actions. The control of the free asymptomatic subpopulation is essential, since this subpopulation contributes to the largest number of patients and, therefore, characterizes the greatest spread of this disease.

It is known that the results obtained when solving the problem of estimating the optimal parameters depend on the choice of the model, which in turn gives feedback to the modelling process. Therefore, any result is not definitive, but rather contributes to simulate a behaviour close to reality.

The sensitivity of the non-linear optimization methods with respect to the choice of the initial points, in their convergence to local minima, was the reason for the hybridization of the Simulated Annealing metaheuristics with classical Quasi-newton-type methods to solve the parameter estimation

problem. It was programmed in MATLAB language version R2018a with the use of units such as *fmincom* and *fminsearch*.

The investigations carried out and the results obtained are still partial and preliminary and will require strategies to refine the data set, which in the present regrowth are not particularly smooth. Therefore, the use of statistical techniques and data analysis is required to provide definitive conclusions. However, what is presented here has an intrinsic value, since it has contributed and still contributes to provide various and multiple tools for decision-makers understand and control epidemic processes, in particular the COVID-19 pandemic caused by the SarsCov2 virus, for the sake of national public health in Cuba and in the world.

References

Chen, T., et al. 2020. "A mathematical model for simulating the phase-base transmissibility of a novel coronavirus." *Infectious Diseases of Poverty.* https://doi.org/10.1186/s40249-020-00640-3.

Elsgoltz, L. 1969. *Ecuaciones Diferenciales y Cálculo Variacional* . MIR.

Gutiérrez, J.M., et J.L. Varona. 2020. "Análisis del Covid-19 por medio de un modelo SEIR." *Blog del Instituto de Matemáticas de la Universidad de Sevilla.*

Lin, Q., et et al. 2020. "A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in Wuhan, China with individual reaction and governmental action." *International Journal of Infectious Diseases,* home page: www.elsevier.com/locate/ijid.

López , R. 2006. "Structure SI Epidemic Models with Applications to HIV Epidemic." *Arizona State University,* 27-46.

Marrero, A., D. Menció, et G. Bayolo. 2020. "Modelo SEAIR con percepción de riesgo para la CoVid19. Caso Cuba." *Revista Ciencias Matemáticas, Vol. 34, No. 1,* 13-18, ISSN: 0256-5374.

Menció, D., G. Bayolo, et A. Marrero. 2020. "Evolución de la CoVid19 a partit de un modelo SAIRV con tasa de transmisión variable ante percepción de riesgo, cuarentena y hospitalización. Caso Cuba." *Revista Ciencias Matemáticas, Vol. 34, No. 1,* 67-72, ISSN: 0256-5374.

Tang, B., et et al. 2020. "An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov)." *Infectious Diseases Modelling* 248-255, Journal hompage: www.keaipublishing.com/idm.

Wu, J.T., K. Leung, et G.M. Leung. 2020. "Nowcasting and forecasting the potential domestic and international spread of the 2019-nCoV outbreak

originating in Wuhan, China: a modeling study ." *Published Online* https://doi.org/10.1016/S0140-6736(20)30260-9.

CHAPTER II

USE OF AGENT-BASED MODELLING FOR DECISION-MAKING IN PANDEMIC CONTROL: COVID-19 CONTEXT. APRIL 2021

MANUEL IGNACIO BALAGUERA*
ANDERSON QUINTERO*
PAULA DANIELA SÁNCHEZ ORTIZ*
AND MERCEDES GAITÁN ANGULO*

Abstract

The available indicators about the way in which the pandemic has evolved in different countries allow us to infer a great diversity of results. Without going too deep, some outstanding scenarios are observed, for which the effects produced by these regulations are distributed in a wide spectrum: at opposite ends, countries like Cuba have established strict measures with strong sanctions, while in countries like Brazil the established measures were lax with low compliance control. In the intermediate zone, there has been a diversity of nuances in the state regulation imposed over the citizen's behaviour regarding the restriction of individual freedoms and of various economic activities.

To date, the results show in some cases a strong impact of the measures on the dynamics of the pandemic, while in other cases this impact has been almost nil (for example, Colombia and Mexico). This reflection strongly motivates a revision of both: employed techniques and the theoretical paradigms from which the different regulations are formulated. For

* Konrad Lorenz Foundation, Bogotá, Colombia.

example, it is observed that in a country like Colombia, although strong control measures have been formulated, there has been a lack of supervision and also of strong sanctions, leading to poor results in terms of the main indicators, placing Colombia among the Latin American countries hardest hit by the pandemic.

In Latin American countries, public policies have been formulated with the support of mathematical models derived from general principles, such as the law of the dynamic equilibrium of markets and the unlimited rationality of both social and economic agents, assuming therefore that all the processes that determine their socioeconomic evolution flows smoothly and continuously. That without considering the occurrence of contingencies and emergent events, which have imposed abrupt and poorly planned changes at all levels of social action to adjust their strategies and action plans in the midst of total uncertainty.

Therefore, the need for a deep rethinking about the epistemologies involved and about the methodologies that support and give scientific basis to the decision-making processes in real time is urgent.

Consequently, this article proposes, discusses and illustrates Agent-Based Modelling as a versatile and flexible alternative to handle complexity and its consequent uncertainty in decision-making processes.

1. Introduction

Throughout history, humanity has faced various pandemics that have impacted the economy, health systems, beliefs and the perception of the reality that surrounds us (León, Rincón and Duque 2020).

Depending on the circumstances of the environment and the factors that affect daily life, the human being adapts mechanisms for decision-making which, depending on the culture, work from the principles that together shape the rationality (Weigand 2006).

According to the above, the rationality of individuals, in a community modulated by the regulatory mechanisms established by governments and other institutions that impact people's lives, plays a fundamental role in the dynamics of the spread of infections during a pandemic. Thus, in the current circumstance caused by the COVID-19 pandemic, a wide spectrum of

contagion dynamics is observed. But what is the result of the interaction between regulation and individual rational response?

To understand how individual rationality adapts to the circumstances of a pandemic, authors such as Herbert Simón (Barros 2010) indicate that it is necessary to understand that crises require developing new skills and making decisions within limited ranges of time and information. The decision-making process is made up of three stages:

1. Intelligence activity
2. Design activity
3. Choice or option activity.

The crucial characteristic of rationality in performance is adaptability, which Gell-Mann (Gell-Mann 1994) addresses with the term "complex adaptive system", for a system that is complex and capable of change and adaptation. In this sense, human beings are considered complex adaptive systems endowed with the extraordinary ability to orient themselves to complex environments.

Authors such as Arce-Rojas (Arce-Rojas 2020) have worked on the approach of complex adaptive systems for the extraction of guidelines for addressing future pandemics through the recognition of the close relationship between nature and society. The author indicates that the adoption of the approach of complex adaptive systems allows us to assume the paradigm of complexity in tackling pandemics by reinforcing reductionist systems approaches.

Authors such as Betancourt et al. (Betancourt, et al. 2009) indicate that decision-making uses advanced scientific methods typical of the complex systems approach, as the study of epidemiological processes is not linear and, therefore, these systems are required.

2. Complexity, complex systems

In the 1960s, the outstanding economist and philosopher (1974 Nobel Prize in Economics) Friedrich von Hayek argued that "A social system is complex according to his work if all information that is needed to describe the state of the system, at any point in time, cannot be collected at one central point" (Hoogduin 2016).

At present, there is still no agreement on a definition of what a complex system is, however there is an agreement as to a set of essential features that characterize a complex system (Lewin 2000). First, the diversity of components and the interactions that connect them. Second, nonlinearity: the fact that the variables characterizing the attributes and behaviours of a complex system cannot be obtained from its simple addition, but they are unknown, forcing the researcher to explore through simulation a diversity of possible interaction scenarios, contrasting the results with features of the dynamics observed in the real systems in order to make the best possible decisions. Third, emergence: as a consequence of nonlinearity, the assemblage of a new system from two or more simpler systems may produce, by synergy, new unexpected categories of objects, properties and behaviours. Finally, contingency: the impossibility to predict the future behaviour of a system only from its initial conditions and constraints, which today is studied under the names of chaos and bifurcations.

As a result of their unique features: diversity of components, nonlinearity, emergence and contingency, complex systems cannot be globally modelled by an enumerable set of deterministic laws expressed by mathematical functions, therefore, the idea of a single set of governing equations (algebraic, differential, integral or a combination of them) is not satisfactory for complex systems modelling. From the early stage of complex systems science, statistical physics has been used as an alternative, establishing the probability associated to a system's state as a function of its associated energy, which is called "ergodic hypothesis". However, including some non-living systems such as "spin glasses", they are non-ergodic (Balaguera, Guzmán and Díaz 2000): they present states with higher probability in comparison to states associated to the same energy level. A consequence of non-ergodicity, the main theoretical tool of statistical mechanics, the partition function of the system, becomes highly limited in order to realize predictions and prognostics. In the case of a system having autonomous agents as components, such as cells in biological tissue or people and human organizations in social systems ("fabrics"), their autonomy produces an ergodicity breaking, making inappropriate the hypothesis that they may be considered as identical particles and consequently to associate them a partition function.

3. Object-oriented and agent-based modelling of complex systems

From the classical perspective of operations research, the prevailing idea of a model is that of a mathematical model: a representation built in terms of selected variables and equations and built around the processes, as well as "the principles that govern the system" conceiving the system under study as a flat structure without hierarchical depth. The main idea underlying this perspective is that of a centrally controlled system which always works by obeying the same dynamic rhythms so that there is temporal regularity and spatial (structural) homogeneity. This conception is the antipode of self-organization, adaptability and autonomy that characterize complex systems.

As an alternative to the classic approach to operations research, focused on the omnipotence of a system of equations, object-oriented methods appeared in the 1980s, first as a solution to software development difficulties, consisting of the need to develop reusable, adaptable and modular solutions that could evolve as fast as the systems for which they intended to be a solution change (Booch, et al. 2007). Many researchers progressively found in the object-oriented methods an ideal epistemological alternative for complex systems modelling, clearly and definitively separating the modelling and simulation activities: one thing is to develop a model: "to model"; another is to manipulate it in a controlled manner in a given scenario: "to simulate".

After the consolidation in the 1990s of the object-oriented programming as the standard methodology for software development, the UML (Unified Modelling Language) language emerged as a visual language whose different diagrams (class, use case, state transition, interaction, and approximately 10 more diagrams) allow to abstract and represent the "knowledge planes" of a complex system under study (Rhem 2006). UML has currently become the standard for visual modelling of complex systems.

The deep epistemological impact of the object-oriented methods and their suitability for the modelling of complex systems has been overshadowed by the misleading idea that they are the same thing as the object-oriented programming, OOP. Despite the importance of the OOP as a paradigm for not "reinventing the wheel" in software engineering, and its contributions to parallel programming, the OOP may be surpassed as programming methodology by other methodologies such as the functional programming or other hybrid ones. In opposition to the OOP as a tool suited for machine

understanding and information processing efficiency, the object-oriented methods, OOM, such as the object-oriented analysis and the design are ideal for human understandability, communication and, mainly, for transdisciplinary research knowledge synthesis. A very common feature of the complex systems is their hierarchical nature, both in structure (composition relations) as in functional dynamics (interaction relations).

Regarding complex systems modelling, agent-based modelling has emerged as a complementary, natural evolution to the object-oriented methods, the Agent Based Modelling. Considered as part of the Artificial Intelligence methods, the Agent-based Modelling is used in a high diversity of fields of study dealing with complex systems: smart materials design, engineering: biomedical, environmental, industrial, and, of course, all the knowledge fields related with social systems and human organizations (Cioffi-Revilla 2014) (López-Paredes 2004).

In a given universe of discourse and for a given scenario, both, objects and agents, have predefined behaviours, computationally implemented by the way of "methods": external methods which activate as a response to an external signal ("a call") or internally as part of a regulatory internal process (homeostasis) or a task which executes as part of a response to an external call. We can affirm that the class "agent" is a category of specialized objects ("subclass"), so that an agent, besides being a reactive object, has autonomy, thanks to which it executes processes (methods) without the need of an external call, with the purpose of adapting to an external environment (a state space) to which it "permanently" monitors and thus maintain a predefined meta-stable state or achieve some kind of optimal state. The authors suggest calling "rationality engine" of an agent to the system which controls its state.

At the present, there are at least 20 Agent Based Modelling platforms, many of them freely distributable. NetLogo (Wilensky 1999) is the most widely used ABM platform: it has a very friendly user interface, excellent user and learning support: its learning curve is very flat. There is a huge number of scientific papers based on NetLogo models and simulation and, which is especially important, it is freely offered to the scientific computing community.

NetLogo's visual output may be a 2D or 3D square or cubic grid of "patches", which represents, in addition to the physical Euclidean space, any 2D or 3D state space. Each patch may represent either a static agent or

the state of a dynamic agent ("turtle" in NetLogo's jargon). The modeller can assign any desired and assumed meaning to the state space: customer satisfaction, income level, temperature, etc. REPAST SIMPHONY (North, et al. 2013) is a high end, powerful alternative to NetLogo, being also free. REPAST SIMPHONY models may be implemented in a wide variety of computer languages such as Java, C#, C++, Python, and LOGO.

In addition to allow HPC, REPAST implements an extraordinary subjacent mathematical epistemology, offering to modellers with a strong mathematical background a robust set of concepts and representative gadgets such as contexts and projections which facilitates the task of a model of a complex system.

As a closing statement, we maintain that the sciences and the engineering of complexity and the complex systems perfectly fit the universe of operations research discourse, and with them all the powerful epistemological novelties that these sciences entail as well as their tools par excellence: scientific computing, scientific visualization and artificial intelligence.

4. Agent Based Modelling (ABM) as an alternative to SIR models for complex social systems

The SIR model is one of the most used models for the characteristics of epidemiological outbreaks, where the susceptible, infected and recovered population can be detected along with the time of infection of each individual. These models are strictly mathematical, which means that their predictive power works well in dynamics that are sufficiently regular and homogeneous to be written in the form of differential equations (ordinary or partial). However, the dynamics of a pandemic such as COVID-19, due to its economic, geographical and human context, escapes the reality required for the proper functioning of the SIR models, which can be verified with the occurrence of multiple peaks in the majority of the countries (Correoso 2021).

Unlike mathematical models, and at the cost of accuracy, computational models and the one that is best adapted to COVID-19 is the Agent-Based Modelling (ABM), which allows the possibility of modelling complex social systems in which the autonomy of the components and their limited rationality is a source of complexity. The ABM represents the behaviour

and interactions of individual agents, objects that make up the system and thus a pattern of behaviour is obtained in it (Cardoso 2016).

Putting these two models together identifies the agent, its heterogeneity, hierarchies of its components, the interdependence between the actors and space in time. Thus, the susceptibility, the risk of infection and recovery or death of the agents can be identified according to their behaviour, affective in primary emotions, micro-social with behaviour in class 3 and organizational in behaviour class 4. The dynamics of contagion and the process that evolves over time through a simulation since it started, currently realizing the real and ideal prognosis with the relationship of the two models in Latin America.

The agent-based model is used to simulate the individual behaviours that determine the evolution of a system where there is a collection of agents, behaviour rules and an environment. An agent represents a person, group or organization to which it responds by acting according to the rules that correspond to its behaviour towards it. Each agent is different. It defines the relationships between the agents and the additional environment that shows the actions in sequence over time (LANCIS 2021).

The modelling and computational simulation based on agents can help in social sciences to study complex phenomena. Since through artificial societies, it allows to study the emergence of phenomena; these allow us to study behaviours at a macro level through micro behaviours. Agent-based simulation can build a production system. What stands out in this approach is that it has mechanisms to have information about the environment and this is stored and can be used in other action (Vélez-Torres 2019).

5. Mechanisms of Contagion

According to the Centre for Disease Control and Prevention, COVID-19 spreads through close contact between people, if they are physically closer than six feet. COVID-19 is most commonly spread during close contact. People who are infected but have no symptoms can also spread the virus to other people, and reinfection cases have been found. The virus spreads more efficiently than influenza, but not as effectively as measles, which is one of the most contagious viruses. When a person sneezes, talks or coughs, small and large respiratory droplets can form particles by drying very quickly in air currents, causing the virus to spread more easily. Respiratory drops cause

infections when inhaled or deposited on mucous membranes, such as those that line the inside of the nose and mouth.

Sometimes it can spread through airborne transmission, which is one of the main ways of spreading infections such as tuberculosis, measles and chickenpox. Some infections can spread through exposure to the virus present in the small respiratory particles and droplets that remain in the air for minutes or hours. They can infect people who are more than 6 feet away from the infected person or after the person has left the place where there is evidence that this has occurred. COVID-19 spreads less frequently through contact with contaminated surfaces, respiratory droplets that can fall on a surface or object, and someone who has contact with them may be infected by the virus, but it's not a common way of spreading it (CDC Center for Disease Control USA 2021).

The World Health Organization confirms that respiratory infections can be transmitted through respiratory droplets and droplet nuclei. This means that COVID-19 is transmitted mainly between people through contact and respiratory droplets. In an analysis that was carried out in China, no airborne transmission was reported, and they affirmed that contagion occurs through close contact less than a meter away from an infected person. The most common symptoms are respiratory, such as coughing or sneezing, since the risk of the mucous membranes of the mouth, nose and eyes to be exposed to droplets that can be infectious.

COVID-19 can be directly transmitted by an infected person or indirectly by contact with surfaces that are in the immediate environment. Other forms of contagion have been evaluated, and the second most probable is contagion by air. A study has also been done in which this virus was cultured from a single stool sample, since COVID-19 can cause intestinal infection and be present in faeces (CDC Center for Disease Control USA 2021).

The protocols that are being used to prevent the spread of the virus are: the use of a mask that covers the nose and mouth to protect oneself and other people; keep one meter (6 feet) distance from outside people; receive the vaccine against COVID-19; constantly wash hands with soap and water; stay isolated (if you have any symptoms) and clean the surfaces with which you are frequently in contact.

6. Methodology

In order to illustrate the use of agent-based modelling to explore and estimate the impact that the establishment of rules for the social behaviour of a community produces in the dynamics of an epidemic, an application was developed in NetLogo that allows the construction of a model of city from its basic demographic information: size, population, inhabitants per house, number of closed public areas and amount of free area. Figure 1 illustrates the application's input interface.

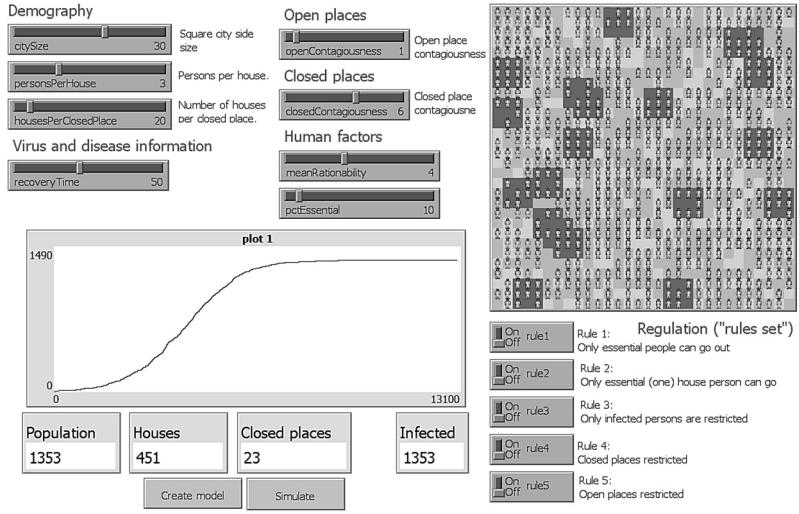


Fig 1: Panoramic view of NetLogo modeling framework.
Source: screen capture from author's development in NetLogo© (Wilensky 1999)

The purpose of this chapter is to provide the reader with a basic methodology that allows the development of a software framework that will serve to create agent-based models from which to run simulation experiments in various scenarios. For this reason, we concentrate here on describing the methodology of its implementation and the use of two fundamental tools: A UML diagram editor (mainly class diagrams) and an IDE (Integrated Development Environment), in this case NetLogo.

Figure 2 shows the class diagram corresponding to an object-oriented model of a city including basic infrastructure, population and social organizations described in a generic way.

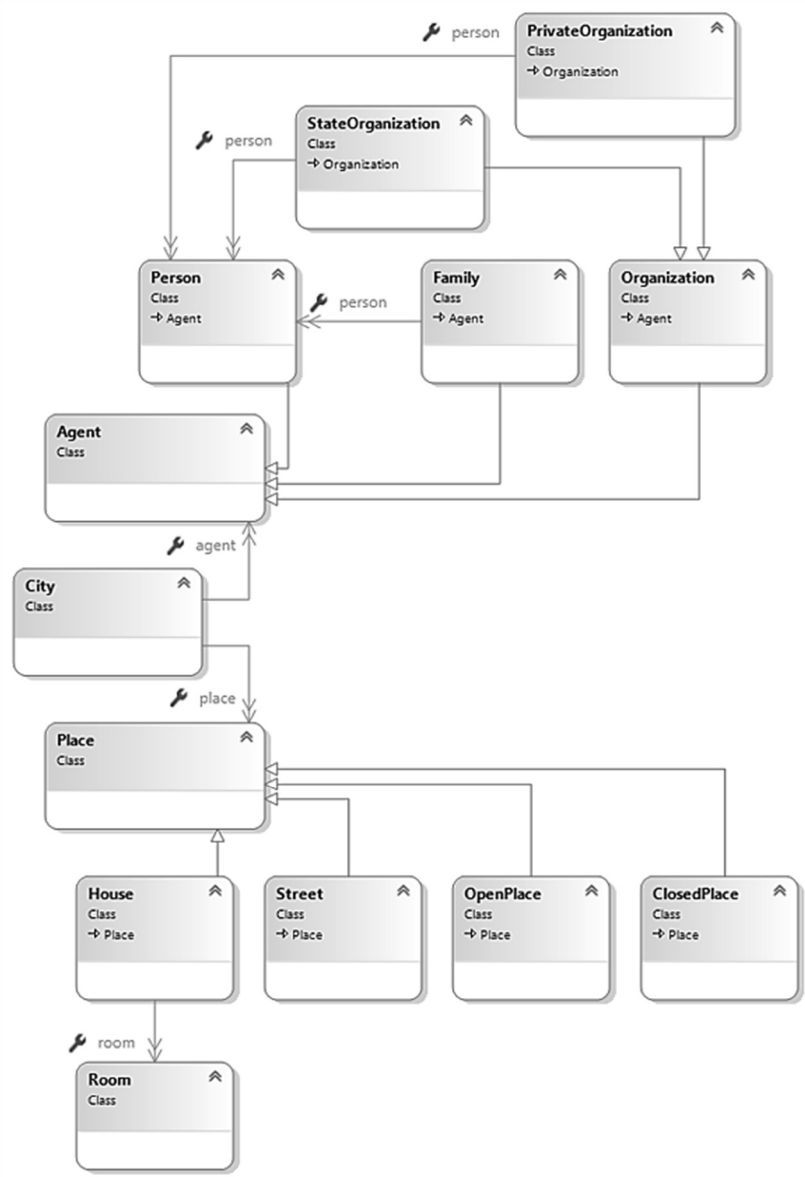


Fig 2: class diagram of a city in the context of an ABM epidemy model.
Source: screen capture from author's development in Visual Studio© class designer.

The diagram represents the classes present in a model and their relationships. A class is a category of objects with a common identity description in terms of attributes and behaviours. From a single class definition, it is possible to generate any quantity (limited by the computer memory) of "class instances" or simply, "objects" individually distinguished from other objects of the same class by the actual values of their defining attributes.

The arrows connecting classes in the class diagram specify the type of association between two classes. Open arrows indicate "composition association". A single open arrow is used if only one lower-level object is a component of the higher class, for example, one "occupation system" includes only one "environment". Double open arrows indicates that a higher-level object may be composed by a "field" of lower-level objects, understanding by "field" a connected array of objects.

As an example, in the class diagram shown in Figure 2, the classes Person, Family and Organization are inherited classes from the "Agent" Superclass. In the same way, class "State Organization" and class "Private Organization" are inherited from their "organization" Superclass.

On the other hand, lower part of the class diagram shows compositional (is-part-of) relations between the infrastructure components of a city, linked between by closed tip arrows ("composition association").

As an example, figures 3 and 4 illustrate the internal structure of a class: attributes of the "Person" class and its behaviour "move" in Figure 2, attributes of class "City" in Figure 4.

Fig. 3. Internal structure of the class "Person".
Source: screen capture from author's development in Visual Studio© class designer.

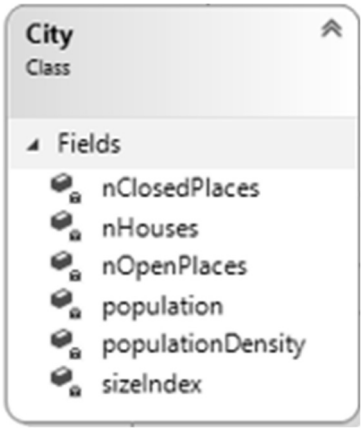


Fig. 4. Internal structure of the class "Person".
Source: screen capture from author's development in Visual Studio© class designer.

7. Results

Figure 5 shows the main interface of the developed application. In the upper left are the sliders blocks: Demography, regulation, parameters of open places, parameters of closed places, human factors and information on viruses and disease. It should be noted that in the regulation block, instead of sliders there are switch buttons that activate for the simulation to execute rules 1 to 5: “Only essential personnel can leave” (rule 1), “Only one person per house can leave” (rule 2), “Only infected people are restricted” (rule 3), “Restricted closed places” (rule 4), finally “Restricted open places” (rule 5).

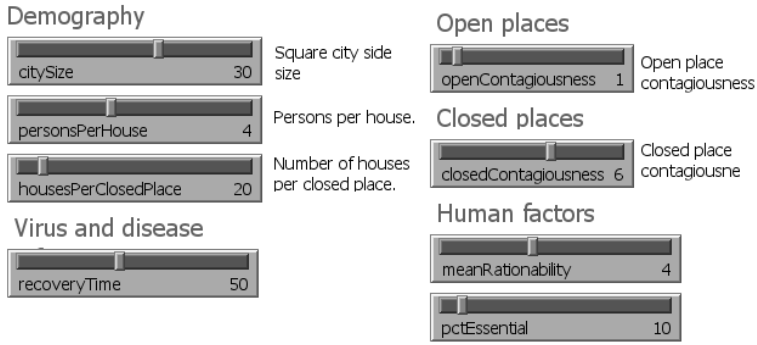


Fig. 5: Input interface of modelling framework.
Source: screen capture from author’s development in NetLogo©.

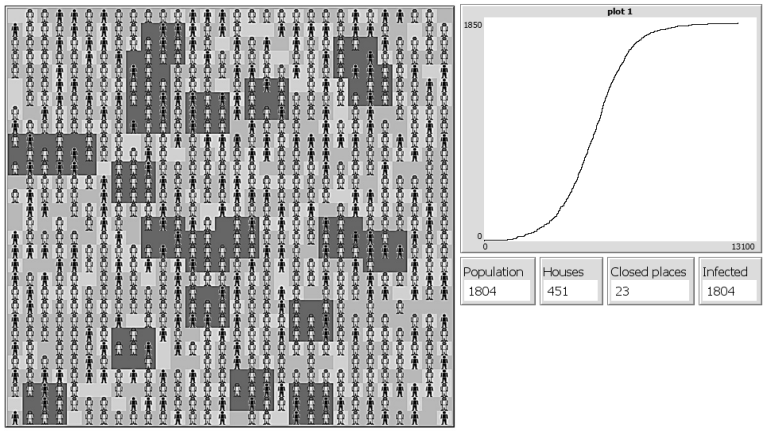


Fig. 6 output interface of modelling framework.
Source: screen capture from author’s development in NetLogo©.

Figure 6 shows the simulation result obtained in a verification exercise in which the graph illustrates the contagion dynamics without any imposed restriction or consideration of the recovered cases. This is to give space to future works whose treatment and analysis are outside the scope of this document.

8. Conclusions

The object-oriented methodology for the construction of models based on epidemic agents and their visual representation through UML class diagrams was illustrated.

The structure and operation of an application in NetLogo developed from class diagrams for the simulation of contagion dynamics in an epidemic was illustrated.

It is expected that this software, released by the community for free access, will serve as a starting point for more specific later developments and applicability to specific scenarios.

Bibliography

Arce-Rojas, R. Relaciones naturaleza y pandemia desde la perspectiva de los sistemas complejos adaptativos. PLURIVERSIDAD, n° 6, 13-31, 2020.

Balaguera, M. I., Guzmán, O., and Díaz, J. M. Fenómenos de Relajación en Vidrios de Espín Bidimensionales. Revista Colombiana de Física, 1, n.° 32, 125-130, 2000.

Barros, G. Herbert a. Simon and the concept of rationality: Boundaries and procedures. Brazilian Journal of Political Economy, 30, n° 119 455-472, 2010.

Betancourt, J., Ortíz, E., González, A., and Brito, H. Enfoque de los sistemas complejos en la Epidemiología. Revista Archivo Médico de Camagüey, 4 n.° 13, 2009.

Booch, G., Maksimchuk, R., Engle, M., Young, B., Conallen, J. and Houston, K. Object-Oriented Analysis and Design with Applications. Boston MA USA: Addison Wesley, 2007.

Cardoso, C. Modelos Basados en Agentes (MBA): definición, alcances y limitaciones. La Plata, Argentina: International Institute for Global Change Research, 2016.

CDC Center for Disease Control USA. Cómo se propaga el COVID-19. CDC - COVID-19, 2021. https://espanol.cdc.gov/coronavirus/2019-ncov/prevent-getting-sick/how-covid-spreads.html.

Cioffi-Revilla, C. Introduction to Computational Social Science. London: Springer Verlag, 2014.

Correoso, K. Modelo SIR, modelo epidemiológico, con Python, 2021. https://pybonacci.org/2020/09/16/modelo-sir-modelo-epidemiologico-con-python/

Gell-Mann, M. The Quark and the Jaguar. Adventures in the Simple and the Complex. Oxford, U.K.: Abacus, 1994.

Hoogduin, L. Decision making in a complex and uncertain world. Future Learn: Decision making in a complex and uncertain world, 2016. https://www.futurelearn.com/courses/complexity-and-uncertainty/0/steps/1831.

LANCIS. Modelación Basada en Agentes. 2021, http://lancis.ecologia.unam.mx/iai/modelacion_agentes.

León, V. E., Rincón, E. E., and Duque, L. Revisión y análisis de las pandemias más devastadoras de la humanidad: de la antigüedad hasta la actualidad. NURE Investigación, 17, n.° 108, 2020: 1-15.

Lewin, R. Complexity, Life at the Edge of Chaos. Chicago: University of Chicago Press, 2000.

López-Paredes, A. Ingeniería de Sistemas Sociales. Valladolid, SPAIN: Universidad de Valladolid, 2004.

North, M. J., Collier, N. T., Ozik, J., Tatara, E., and al., e. Complex adaptive systems modelling with Repast Simphony. Complex Adaptive Systems Modelling, 1, n.° 3, 2013.

Rhem, A. J. UML for Developing Knowledge Management Systems. London, U.K.: Auerbach publications, 2006.

Vélez-Torres, Á. Modelación y simulación basada en agentes en ciencias sociales: una aproximación al estado del arte. POLIS Revista Latinoamericana, 53, 2019: 1-22.

Weigand, E. Argumentation: The Mixed Game. Argumentation, 20, 2006: 59–87.

Wilensky, U. NetLogo. 1999, http://ccl.northwestern.edu/netlogo/.

CHAPTER III

QUANTITATIVE CHARACTERIZATION OF COVID-19 IN MEXICO

PABLO OTONIEL JUÁREZ MORENO[1]
CARLOS N. BOUZA HERRERA[2]
JUAN MANUEL SÁNCHEZ REBOLLEDO[3]
OCTAVIANO JUÁREZ ROMERO[1]

Abstract

This paper describes the behaviour of COVID-19 in Mexico from the first confirmed case on February 28th, 2020, until May 2021. The behaviour of the pandemic in Mexico has been analysed using the correlation coefficient of three relevant variables of COVID-19 against two indices-summaries. Relevant COVID-19 variables included are incidence, mortality rate and fatality rate. They have been analysed with the Marginalization Index and the Human Development Index. These are measurements on the socioeconomic conditions of the states of Mexico.

Keywords: Human Development Index, Marginalization Index, indices-summaries.

Introduction

The disease caused by the SARS-COV-2 virus, called coronavirus-19 (COVID-19), originated in the city of Wuhan, China, in late 2019. A

[1] Universidad Autónoma de Guerrero, Mexico
[2] Universidad de La Habana, Cuba
[3] Coordinator of Epidemiological Surveillance of the Sanitary Jurisdiction 07 Guerrero, Mexico

characteristic of this disease is its rapid spread among humans through the respiratory tract. For this reason, on January 30th, 2020 a pandemic was declared by the World Health Organization. Due to the impact that the pandemic has caused in different parts of the world and particularly in Mexico, the present work aims to carry out a statistical-descriptive analysis of the open data provided by the Secretaría de Salud of the federal government of Mexico. The database that has been worked on includes the 2,432,280 confirmed cases of COVID-19 in Mexico. These cases are those reported on the official page of the Mexican government for COVID-19 from the start of the pandemic in the country, February 2020 up until May 27th, 2021.

In this Part I, we describe the behaviour of COVID-19 in Mexico from the first confirmed case on February 28th, 2020 until May 2021. This characterization of COVID-19 in Mexico is through positive cases, cases with comorbidities, number of deaths, mortality rate, incidence rate and fatality rate; represented by state, sex and age group.

In part II, to expand the analysis of pandemic behaviour in Mexico, the correlation coefficient of three relevant variables of COVID-19 has been calculated against two indices-summaries, for the data of the states. The relevant COVID-19 variables that have been considered in the calculation are incidence, mortality rate and fatality rate. These have been analysed with the Marginalization Index and the Human Development Index. These are measurements on the socioeconomic conditions of the states in Mexico.

At the beginning of the pandemic at the international level, the Mexican government established different sanitary measures to contain COVID-19 infections. Some of these measures were the suspension of non-essential activities and school activities; the implementation of campaigns to stay at home, hand washing and healthy distance, among others. Despite these containment measures, the number of infections grew beyond what the authorities expected. The purpose of this paper is to characterize the behaviour of the COVID-19 pandemic in Mexico.

Part I. Analysis of the COVID-19 pandemic in Mexico

1. Behaviour of the pandemic over time

The behaviour of positive cases of COVID-19 in Mexico is shown in Figure 1-1. This shows two moments in which the number of cases reached its maximum values, one around August 5th, 2020, with 12,729 cases, and the

other around January 27th, 2021, with 27,944 cases. In the first peak, reaching a daily average of almost 3,858 cases between July 5th and September 5th, 2020. In the second peak, from December 27th to March 27th , 2021 a daily average of 11,185 positive cases was reached. In the period from March 27th to May 27th, 2021, there was a decrease in daily cases, reaching a daily average of almost 3,000 cases. As a special case, a third peak with 14,057 positive cases on October 6th, 2020, is not included, since this increase is an isolated event, as around that date the average of cases remained constant at 5,000 per day, unlike the other two peaks, when the increase and decrease were gradual. The peak on the aforementioned date can be explained by late reports of positive cases and their subsequent capture on the same date.

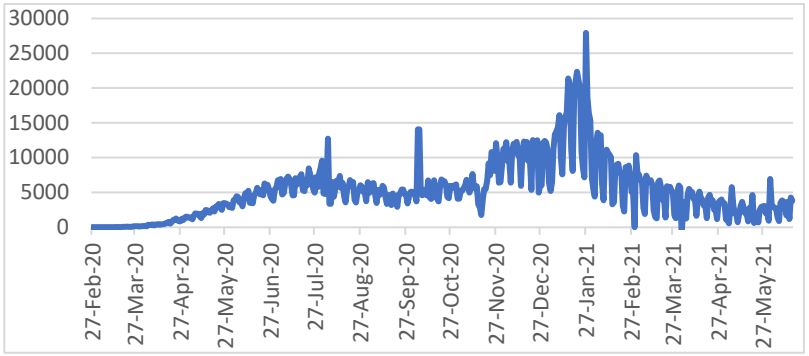


Fig. 3.1-1 Daily number of COVID-19 cases since February 2020 up to May 2021 (Own elaboration)

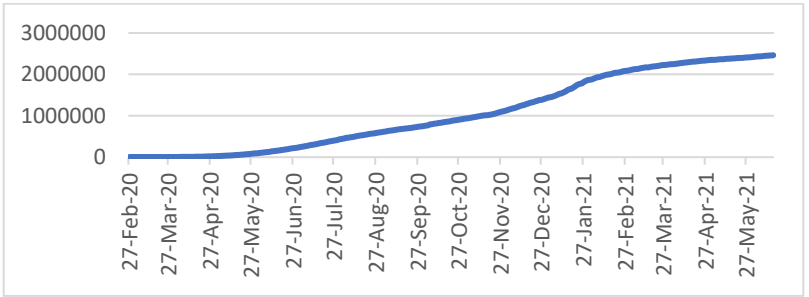


Fig. 3.1-2 Number of accumulated COVID-19 cases since February 2020 up to May 2021 (Own elaboration)

Accumulated cases of COVID-19 since February 28th, 2020, up to May 27th, 2021, are shown in Figure 1-2. As they are accumulated, their behaviour

increases, although with differentiated growth rates in each period. In the case prior to the first peak, that is from the July 5th to August 5th, 2020, the daily average rate was 2.15%, while in the period from the August 5th to September 5th the average growth rate was 1.25%. In the case of the second peak of the pandemic, from December 27th, 2020, to January 27th, 2021, the average rate was 0.83%, and in the period after the peak, from January 27th to February 27th, 2021, the daily average rate was 0.50%. Finally, in the stabilization period of the pandemic, from the March 27th to May 27th, 2021, the daily average rate was 0.13%.

2. Distribution by federal entity

In the previous point, the confirmed and accumulated cases of COVID-19 in Mexico were described in a general way. Now, the behaviour by states of the accumulated cases until May 2021 is analysed, and is found in Table 2-1. The two states that have had the highest number of verified COVID-19 cases in Mexico in absolute terms are Mexico City with 662,631, which corresponds to 27.2% of the total cases, and the Estado de Mexico with 253,272 cases, which corresponds to 10.4% of the total, namely. For 1 confirmed case in the Estado de México, there are 2.6 cases in Mexico City. These confirmed case totals are congruent with the population size of these states, since they occupy the second and the first place of the states with the largest number of inhabitants and with the population density, since Mexico City is in first place and the Estado de México is in second place (Instituto Nacional de Estadística y Geografía 2021, INEGI). The entity with the lowest cases of COVID-19 is the state of Campeche, with 10,373 cases, which represents 0.4% of the total. It is followed in increasing order by the entities of Chiapas, with11,636 cases, Colima, with 12,022, and Nayarit with 12,277 cases. These amounts represent approximately 0.5% of the total cases.

To establish a comparison of pandemic behaviour among the federal entities, it is necessary to consider the population that resides in these entities. For this reason, the incidence is calculated, that is, the number of cases per 100,000 inhabitants and we can express it as:

$$I = \frac{c}{p} * 10^5$$

where, c=number of COVID-19 cases, p = number of inhabitants by federal entity. This data is reported in the last column of Table 2-1. The entities with the highest incidence are Mexico City and Baja California Sur with

7,195 and 4,118, respectively. The entities with the lowest incidence are Chiapas and Nayarit with 210 and 994 cases, respectively. The national incidence data is 1,930 cases per 100,000 inhabitants.

Federal Entity	Frequency	Percentage	Population	Incidence
Aguascalientes	26,570	1.1	1,425,607	1,863.77
Baja California	49,710	2.0	3,769,020	1,318.91
Baja California sur	32,879	1.4	798,447	4,117.87
Campeche	10,373	0.4	928,363	1,117.34
Coahuila	68,990	2.8	3,146,771	2,192.41
Colima	12,022	0.5	731,391	1,643.72
Chiapas	11,636	0.5	5,543,828	209.89
Chihuahua	56,479	2.3	3,741,869	1,509.38
Ciudad de México	662,631	27.2	9,209,944	7,194.73
Durango	34,311	1.4	1,832,650	1,872.21
Guanajuato	132,160	5.4	6,166,934	2,143.04
Guerrero	41,273	1.7	3,540,685	1,165.68
Hidalgo	39,098	1.6	3,082,841	1,268.25
Jalisco	87,443	3.6	8,348,151	1,047.45
México	253,272	10.4	16,992,418	1,490.50
Michoacán	48,869	2.0	4,748,846	1,029.07
Morelos	33,843	1.4	1,971,520	1,716.59
Nayarit	12,277	0.5	1,235,456	993.72
Nuevo León	124,626	5.1	5,784,442	2,154.50
Oaxaca	47,432	2.0	4,132,148	1,147.88
Puebla	85,415	3.5	6,583,278	1,297.45
Querétaro	69,300	2.8	2,368,467	2,925.94
Quintana Roo	28,070	1.2	1,857,985	1,510.78
San Luis Potosí	64,498	2.7	2,822,255	2,285.34
Sinaloa	39,604	1.6	3,026,943	1,308.38
Sonora	75,806	3.1	2,944,840	2,574.20
Tabasco	69,510	2.9	2,402,598	2,893.12

Tamaulipas	60,362	2.5	3,527,735	1,711.07
Tlaxcala	19,985	0.8	1,342,977	1,488.11
Veracruz	62,064	2.6	8,062,579	769.78
Yucatán	40,944	1.7	2,320,898	1,764.14
Zacatecas	30,828	1.3	1,622,138	1,900.45
Totals	2,432,280	100.0	126,014,024	1,930.17

Table 3.2-1. COVID-19 cases distribution and incidence, according to the federal entity (Own elaboration)

3. Distribution by age groups

The 2020 Population and Housing Census of Mexico (Consejo Nacional de Población, Gobierno 2021), counts 126,014,024 inhabitants of which 51.2% corresponds to women and 48.8% to men, with a median age of 29 years. The pyramidal structure of the Mexican population has a higher percentage at its base, 25.2% for the population from 0 to 14 years old. The next group from 15 to 24 years old represents 16.9%. The following groups of 10 years decrease the percentage until the group of 75 to 84 years with 2.4% and those over 85 years old represent 0.8% of the total population (see Figure 3-1). Next, the age groups that are most affected by COVID-19 are analysed.

Table 3-1 shows the distribution of COVID-19 cases by age and sex groups. The group from 25 to 34 years old is the one that concentrates the highest number of cases, with 521,789, that represents 21.45% of the total, followed in decreasing order by the group from 35 to 44 years old, with 497,853 cases, representing 20.47%, and in third place is the group of 45 to 54 years, with 475,143 cases, that represents 19.53%. These three groups together represent 61.45% of the total COVID-19 cases in Mexico. The group from 0 to 4 years old concentrates 12,059 cases, representing 0.5%, followed in increasing order by the group aged 85 years and over with 26,853 cases, representing 1.1% of the total.

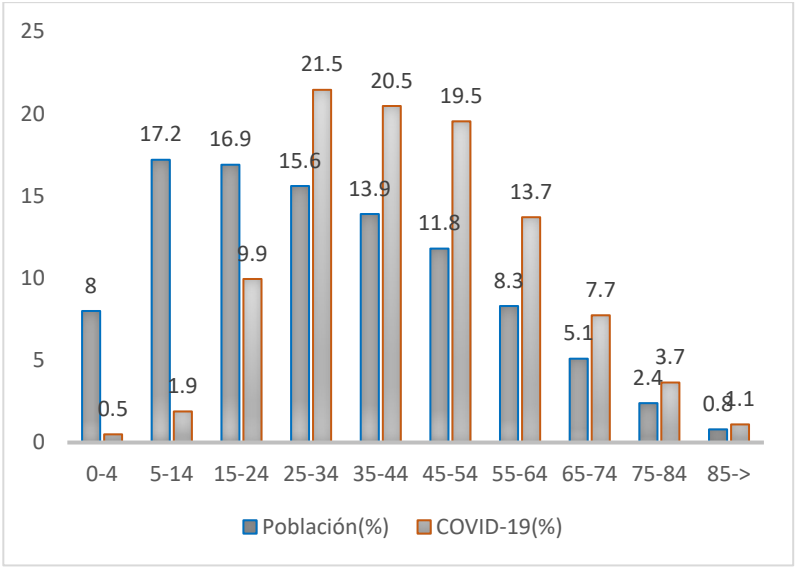


Fig. 3.3-1. Comparison of the population (%) and COVID-19 cases (%), by age groups (Own elaboration)

In the distribution of COVID-19 cases by sex, the group of women concentrates 1,214,981 cases, representing 49.95%, and men 1,217,299 cases, with 50.05%. In the case of age groups of women, the same situation is reproduced in the general case, with the percentages slightly varying. In the case of age groups for men, exactly the same thing happens.

Age groups	General		Women		Men	
	Absolut	%	Absolut	%	Absolut	%
0-4	12,059	0.50%	5,646	0.46%	6,413	0.53%
5-14	45,989	1.89%	22,628	1.86%	23,361	1.92%
15-24	241,978	9.95%	125,611	10.34%	116,367	9.56%
25-34	521,789	21.45%	265,777	21.87%	256,012	21.03%
35-44	497,853	20.47%	250,907	20.65%	246,946	20.29%
45-54	475,143	19.53%	240,752	19.82%	234,391	19.26%
55-64	333,442	13.71%	162,753	13.40%	170,689	14.02%
65-74	188,298	7.74%	87,351	7.19%	100,947	8.29%

75-84	88,876	3.65%	40,670	3.35%	48,206	3.96%
85->	26,853	1.10%	12,886	1.06%	13,967	1.15%
Totals	2,432,280	100.00 %	1,214,98 1	100.00 %	1,217,29 9	100.00 %

Table 3.3-1 Distribution of COVID-19 cases in Mexico, by age groups (Own elaboration)

4. COVID-19 and comorbidities more frequently in Mexicans

For decades, the health conditions of Mexicans have been worsening, which is the result of poor nutrition or consumption of unhealthy foods, insufficient access to health services, and working periods of more than 12 hours a day, among other factors. All this has led to a decrease or complications in personal health. This is reflected in chronic diseases such as diabetes mellitus, hypertension and obesity (Encuesta Nacional de Salud y Nutrición 2021). Consequently, in Mexico 8.6 million people suffer from diabetes mellitus, 15 million suffer from arterial hypertension and 72.5% of Mexicans are overweight or obese, which is equivalent to 91.3 million people. If the person has at least one of these diseases and is also positive for COVID-19, this leads to a case of comorbidity.

For the aforementioned, as it is important to know the cases of COVID-19 and their comorbidities, the characteristics of these comorbidities are presented in Table A-1. The comorbidity with the highest percentage is hypertension, with 17.1%, that is, 415,001 people of the total 2,432,280 positive cases. In the second place is pneumonia, with 345,747 cases, which represents 14.2% of the total; in the third place is obesity, with 343,877, which represents 14.1% of the total; in the fourth place is diabetes, with 319,374 cases, that represents 13.1% of the total. Therefore, at least 58.5% of the positive cases for COVID-19 presented one of the above comorbidities.

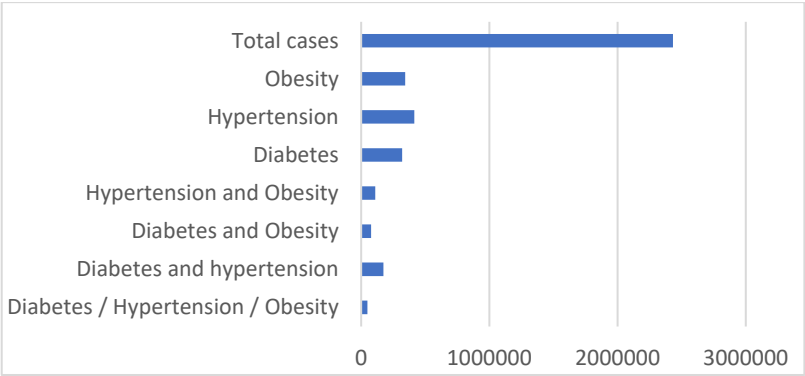


Fig. 3.4.1 Distribution of the most frequent comorbidities with COVID-19 (Own elaboration)

To complete the analysis of the most common comorbidities in Mexico, Figure 4-1 and Table A-2 present the main comorbidities and their distribution by sex. People who presented three comorbidities such as diabetes, obesity and hypertension were 48,930, with 2.01% of the total of COVID-19 cases. This total distributed by sex corresponds to 57.2% to women and 42.8% to men. People who presented the combination of diabetes and hypertension were 173,815, that is 7.15% of the total, being 51.4% women and 48.6% men. In the case of hypertension and obesity comorbidities together, 109,708 cases have been registered with a percentage of 4.51%, being 54.5% women and 45.5% men; and in the case of diabetes and obesity comorbidities together, there are 77,010 cases, representing 3.17% of the total, by sex 55.9% women and 44.1% men. With this, we know that 72.17% of the COVID-19 positive patients had at least some of the comorbidities of diabetes mellitus, high blood pressure, obesity or pneumonia, or, in the worst case, three comorbidities at the same time.

The main comorbidity distributed with respect to sex are very similar, considering the general data. For example, hypertension cases represent 17.06% with respect to the total, the percentage corresponding to women is 17.31% and men 16.82% with respect to the total of women and men respectively. The distribution by sex of the comorbidities of diabetes and obesity has a similar behaviour.

5. COVID-19 deaths by state

One of the most discussed topics about the pandemic in Mexico has been the number of deaths caused by COVID-19. At the time of preparing this paper, Mexico ranks the 4th worldwide position for deaths from COVID-19. To understand the high number of deaths, among other reasons (see Academia.edu 2020), note that the impact of a disease or virus on a person depends on different factors, such as the individual growing up process and evolution, food, treatments, among others. This can be extended to countries or regions (if they are territorially extensive countries), since we are characterized by a certain idiosyncrasy as inhabitants of a country.

Some of the factors that explain the high number of deaths from COVID-19 in Mexico is the size of its population, which is 127.6 million, therefore, deaths must be proportional; 68.61% of the positive cases for COVID-19, that is 1,668,835 people, had some comorbidity, which increased their probability of having health complications. Other factors, as we will see below, are age and sex, since people between 55 and 84 years old represent 69.38% of the deceased, and of the total deceased, 62.49% were men.

Table 5-1 shows the distribution of deaths in Mexico by state and sex, the proportional totals to the population in Mexico and the totals relative to each federative entity. As demonstrated, the entity with the highest number of deaths is the Estado de Mexico, with 37,013; this is 16.18% of the deaths in Mexico. From the total entity, 35.26% corresponds to women and 64.74% to men. Mexico City follows with 34,272 deaths with 14.98%. From these, 36.33% are women and 63.67% are men. In the third place is Jalisco, with 12,456 deaths. This is 5.45 % of the total state, being 37.32% women and 62.68% men. The last places of deaths by entity are Colima, with 1,200 deaths, which corresponds to 0.52% from the total, of which 38.08% are women and 61.92% are men. Follows Campeche, with 1,232 deaths, what is 0.54 % of the total, being 35.31% women and 64.69% men. At the national level, from the 228,754 deaths, 85,796 (37.51%) correspond to women and 142,958 (62.49%) to men.

Federative Entity	Totals proportional to the Population of México		Totals relative to the entity			
			Women		Men	
	Absolute	%	Absolute	%	Absolute	%
Aguascalientes	2,428	1.06	960	39.54	1,468	60.46
Baja California	8,574	3.75	3,432	40.03	5,142	59.97
Baja California sur	1,454	0.64	562	38.65	892	61.35
Campeche	1,232	0.54	435	35.31	797	64.69
Coahuila	6,320	2.76	2,646	41.87	3,674	58.13
Colima	1,200	0.52	457	38.08	743	61.92
Chiapas	1,647	0.72	536	32.54	1,111	67.46
Chihuahua	7,398	3.23	2,997	40.51	4,401	59.49
Ciudad de México	34,272	14.98	12,451	36.33	21,821	63.67
Durango	2,485	1.09	987	39.72	1,498	60.28
Estado de México	37,013	16.18	13,052	35.26	23,961	64.74
Guanajuato	10,973	4.80	4,236	38.60	6,737	61.40
Guerrero	4,490	1.96	1,644	36.61	2,846	63.39
Hidalgo	6,238	2.73	2,167	34.74	4,071	65.26
Jalisco	12,456	5.45	4,648	37.32	7,808	62.68
Michoacán	5,877	2.57	2,197	37.38	3,680	62.62
Morelos	3,465	1.51	1,201	34.66	2,264	65.34

Nayarit	1,848	0.81	678	36.69	1,170	63.31
Nuevo León	9,594	4.19	3,796	39.57	5,798	60.43
Oaxaca	3,779	1.65	1,332	35.25	2,447	64.75
Puebla	11,983	5.24	4,300	35.88	7,683	64.12
Querétaro	4,402	1.92	1,602	36.39	2,800	63.61
Quintana Roo	2,830	1.24	994	35.12	1,836	64.88
San Luis Potosí	5,400	2.36	2,129	39.43	3,271	60.57
Sinaloa	6,241	2.73	2,597	41.61	3,644	58.39
Sonora	6,674	2.92	2,822	42.28	3,852	57.72
Tabasco	4,186	1.83	1,604	38.32	2,582	61.68
Tamaulipas	5,069	2.22	2,038	40.21	3,031	59.79
Tlaxcala	2,503	1.09	924	36.92	1,579	63.08
Veracruz	9,942	4.35	3,723	37.45	6,219	62.55
Yucatán	3,973	1.74	1,518	38.21	2,455	61.79
Zacatecas	2,808	1.23	1,131	40.28	1,677	59.72
Totals	228,754	100.00	85,796	37.51	142,958	62.49

Table 3.5.1 Distribution of deaths from COVID-19 by state and sex (Own elaboration)

To establish a comparison between the entities without the distortion generated by the number of inhabitants, the mortality rate per 100,000 inhabitants is calculated. See Table A-3, where its expression is:

$$M = \frac{d}{p} * 10^5$$

where, d = number of deaths and p = number of inhabitants by federal entity.

In Figure 3.5-1, the highest mortality rate corresponds to Mexico City, with a value of 372, that is, for every 100,000 inhabitants there are 372 deaths associated with COVID-19. This high mortality rate in Mexico City shows it is the second state with the highest number of inhabitants 9,209,944, and the main reason is the population density of Mexico City, which is 6,163 inhabitants per km^2 (Cuentame INEGI 2021). For these arguments, a person is more likely to be infected or to die from COVID-19 if he lives in the City of Mexico, due to the high number of people who inhabit it in small spaces. The following entities are Baja California and Sonora with a rate of 227. In the next position is the Estado de Mexico, with a mortality rate of 218 per 100,000 inhabitants. Although the Estado de Mexico is the entity with the largest population of 16,992,418, its population density is 760, so it does not resemble the situation mentioned above in Mexico City. The last mortality rate is reached by the state of Chiapas, with a value of 30. The penultimate value of the mortality rate is 91 for the state of Oaxaca. The national mortality rate per 100,000 inhabitants in Mexico corresponds to 182.

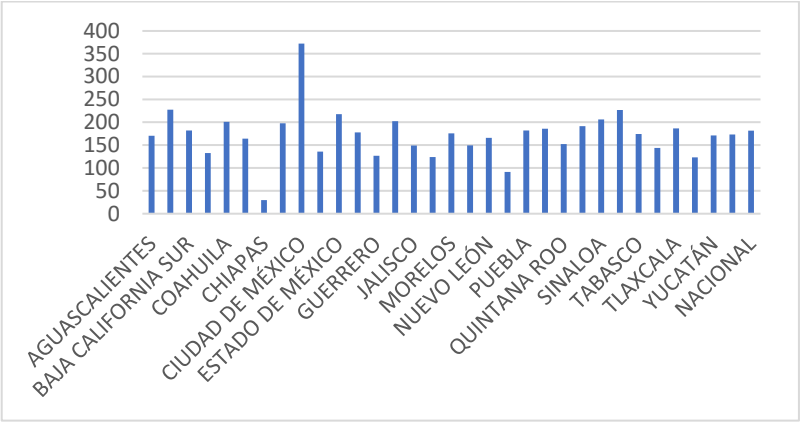


Fig. 3.5.1. Mortality rate per 100,000 inhabitants by state (Own elaboration)

In Table 3.A.3, the last column presents the calculation of lethality, which is defined as the ratio of the number of deaths and COVID-19 cases per 100 (Chávez-Almazán L.A. et al 2021, 2) and can be expressed as:

$$L = \frac{d}{c} * 100$$

where d=number of deaths, c=number of cases COVID-19.

The highest value is the state of Baja California, with 17.2, followed in descending order by the states of Hidalgo and Veracruz, with a value of 16. On the opposite end, Baja California Sur has the smallest value, with 4.4, followed in ascending order by Mexico City, with 5.2. The national lethality figure is 9.4.

6. Deaths caused by COVID-19 by comorbidity, sex and age group

The population that suffers from a chronic disease is the most vulnerable to contracting COVID-19 and its complications. The most prevalent chronic diseases in Mexico are diabetes mellitus, hypertension, obesity, cardiovascular diseases, and kidney failure. According to Hernández-Garduño (2020, 6), obesity represents the strongest predictors of COVID-19, followed by hypertension and diabetes. In Mexico, for the year 2016, 72.5% of adults have problems with being overweight and obesity and in 76.6% of adults there is a prevalence of abdominal obesity and they are a risk factor for suffering from some of the chronic diseases (Campos Nonato et al 2018, 31-40).

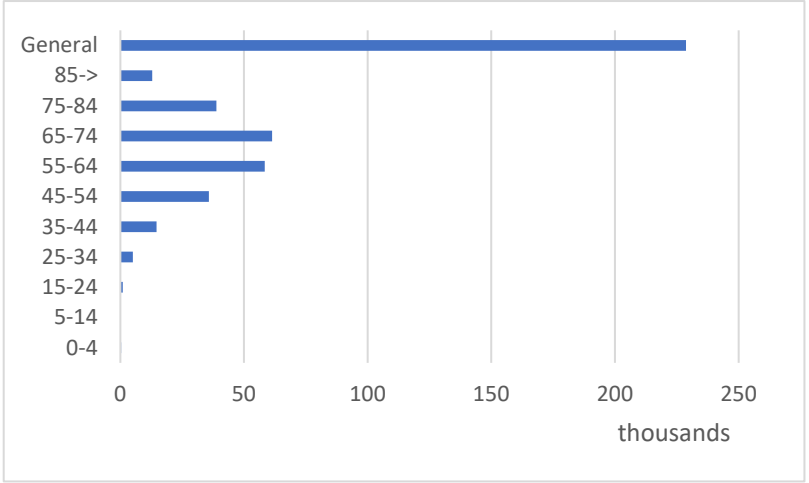


Fig. 3.6.1 Distributions of deaths caused by COVID-19, by age group (Own elaboration)

One of the causes of death relating to COVID-19 can be found in coexistence with other diseases, see Table 3.A.4. The total number of deaths in the study period is 228,754. From these, 102,582 suffered from hypertension, which represents 44.84 of total, being 44,221 (43.11) women and 58,361 (56.89) men. Another disease that is predominant in Mexican people is diabetes, with 84,840 deaths that represent 37.09 of the total, with 35,724 deaths (42.11) corresponding to women and 49,116 deaths (57.89) corresponding to men. Then there is obesity with 49,471 cases, which represents 21.63% of the total, with 21,978 deaths (44.43) for women and 27,493 deaths (55.57) for men. There are cases in which people had more than two conditions or comorbidities, such as hypertension and diabetes with 55,836 cases, which represents 24.41 of all deaths; in the case of hypertension and obesity with 27,362 cases, which represents 11.96 of the total, among other cases observed in the table mentioned above.

From the total number of COVID-19 cases, 763,455 positive people did not report any comorbidity, and of these people, 44,130 died, which is equivalent to a fatality rate of 5.78. In the same way, 1,668,835 people who tested positive for COVID-19 had at least one comorbidity and 184,624 died, so their fatality rate is 11. In other words, having any comorbidity represents almost twice the risk of dying from COVID-19.

Table A-5 and Figure 6-1 show the number of deaths according to the age group they belonged and sex. The age group with the highest number of deaths is from 65 to 74-years-old with 61,403 deaths, which represents 26.84 of the total, being 22,031 (38.99) cases of women and 37,463 (61.01) of men. In descending order, the group from 55 to 64-years-old follows with 58,443 deaths, representing 25.55 of the total. The number by sex is 22,031 (37.70) for women and 36,412 (62.30) for men. For the next group, from 75 to 84-years-old, the number of deaths is 38,862, which represents 16.99 of the total. This quantity distributed by sex corresponds to 15,592 (40.12) to women and 23,270 (59.88) to men. In national terms, of the 228,754 deaths, 85,796 (37.51) correspond to women and 142,958 (62.49) to men.

Part II. Relationship of COVID-19 cases with well-being indices

For this second part of the paper, a statistical analysis is carried out with three of the relevant variables of COVID-19: the incidence (number of cases between number of inhabitants per 10 ^ 5), the mortality rate (number of deaths between number of inhabitants per 10 ^ 5) and lethality (number of

deaths among number of positive cases per 100), against two indicators of the socioeconomic conditions of the states, which are the Marginalization Index and the Human Development Index.

7. Correlation of the Marginalization Index of Mexico with COVID-19 variables

The following table shows the correlations of incidence, mortality, fatality rates, Marginalization Index and The Human Development Index.

Variables	Incidence	Mortality rate	Lethality	MI	HDI
Incidence	1	0.76	-0.68	-0.48	0.70
Mortality rate	0.76	1	-0.18	-0.67	0.75
Lethality	-0.68	-0.18	1	0.21	-0.36
MI	-0.48	-0.67	0.21	1	-0.89
HDI	0.70	0.75	-0.36	-0.89	1

Table 3.7.1 Matrix of correlations of COVID-19 variables with indices (Own elaboration)

The Marginalization Index (Consejo Nacional de Población. Gobierno, 2021) is a measure of the marginalization conditions, in this case of the states, which is calculated as a summary measure of different indicators on education, public services, among others. The higher the index values, the higher the marginalization situation. In this case, the entities with the highest value are the states of Guerrero, Chiapas and Oaxaca; and the entity with the lowest marginalization value is Mexico City. The value of the correlation coefficient between the Marginalization Index and the incidence is - 0.48. This means that when the index values increase, the incidence decreases (perhaps the explanation is due to having few large population centres). The value of the coefficient of correlation of the index with the mortality rate is - 0.67. This means that when the index values increase, the mortality rate decreases. The correlation coefficient of the index with lethality is 0.21. In this case, it means that the higher the marginality, the higher the lethality, although the value of the coefficient is low.

8. Correlation of the Human Development Index with COVID-19 variables

The Human Development Index (HDI) provided by the United Nations Development Programme (El PNUD en México, 2021), aims to measure the development of a geographic area; in this case the states, through three

elements: Gross Domestic Product per inhabitant, health and education (López-Calva et al 2004, 29). The large values of the HDI indicate better conditions, in our case of the states; The highest HDI value is reached by Mexico City, followed in descending order by the state of Sonora; the lowest value is reached by the state of Chiapas, followed by the states of Guerrero and Oaxaca. The value of the correlation coefficient between incidence and HDI is 0.699, which indicates that when the HDI value increases, the incidence of COVID-19 also increases. In the case of the correlation coefficient of the mortality rate and HDI, its value is 0.748. It has the same effect as the previous correlation coefficient. Finally, the correlation coefficient of lethality and HDI has a value of - 0.36. In this case due to the negative sign, when the HDI values increase, the lethality values decrease.

Acknowledgments

The authors give their heartily thanks to the referees´ invaluable comments, which allowed to improve the quality of the early versions of this paper. C. N. Bouza acknowledges the support of CITMA project PN223LH010-005.

Bibliography

"Academia.edu", Academia. Last consulted June 08, 2021. https://bit.ly/3qGxHZn.

Campos Nonato, I., Cuevas Nasu, L., González Castell, L.D., Hernández Barrera, L. 2018.

"Epidemiología de la obesidad y sus principales comorbilidades en México" La obesidad en México. Estado de la política pública y recomendaciones para su prevención y control. 31-40 México: Instituto Nacional de Salud.

Chávez-Almazán LA, Díaz-González L, Rosales-Rivera M. 2021 "COVID-19 y el índice de desarrollo humano en México". Salud Publica Mex. 2 Last consulted June 14, 2021. https://bit.ly/3qNuIyw.

"Consejo Nacional de Población. Gobierno", gob. Last consulted June 20,2021. https://bit.ly/3qEFt65.

"Cuentame INEGI", cuéntame.inegi. Last consulted June 07, 2021. https://bit.ly/3yeZOBv.

Hernández-Garduño, E. 2020. "Obesity is the comorbidity more strongly associated for COVID-19 in Mexico. A case.control study". Elsevier Obesity Research & Clinical Practice, Volume 14, Issue 4, Pages 375.379.

"Instituto Nacional de Estadística y Geografía (INEGI)", inegi. Last consulted June 09, 2021.
https://bit.ly/3hHyBkv.

López.Calva, L.F., Rodríguez-Chamussy L. y Székely, M. 2004. "Medición del Desarrollo Humano en México: Introducción". Estudios sobre desarrollo humano pnud México. No. 2003-6, 29.

"El PNUD en México", mx.undp. Last consulted June 21, 2021. https://bit.ly/3hxfrxE.

"Encuesta Nacional de Salud y Nutrición", ensanut. Last consulted June 11, 2021.
https://bit.ly/3xbmVgf.

Appendix A

Disease	Yes	No
Pneumonia	345,747	2,086,530
%	14.2	85.8
Diabetes	319,374	2,106,838
%	13.1	86.6
COPD	26,545	2,400,117
%	1.1	98.7
Asthma	52,743	2,374,051
%	2.2	97.6
Immunosuppressant	19,848	2,406,761
%	0.8	99
Hypertension	415,001	2,011,653
%	17.1	82.7
Another comorbidity	46,614	2,375,609
%	1.9	97.7
Cardiovascular	37,212	2,389,442
%	1.5	98.2
Obesity	343,877	2,083,124
%	14.1	85.6
Chronic kidney	35,870	2,390,839
%	1.5	98.3
Smoking	177,564	2,248,863
%	7.3	92.5
Another case	1,036,460	1,262,042
%	42.6	51.9

Table 3.A.1. Relationship of positives to COVID-19 with diseases declared

Comorbidity	Total		Women		Men	
	Absolute	%	Absolute	%	Absolute	%
Diabetes / Hypertension / Obesity	48,930	2.01	28,008	2.31	20,922	1.72
Diabetes and hypertension	173,815	7.15	89,399	7.36	84,416	6.93
Diabetes and Obesity	77,010	3.17	43,020	3.54	33,990	2.79
Hypertension and Obesity	109,708	4.51	59,759	4.92	49,949	4.10
Diabetes	319,374	13.13	157,549	12.97	161,825	13.29
Hypertension	415,001	17.06	210,275	17.31	204,726	16.82
Obesity	343,877	14.14	179,730	14.79	164,147	13.48
Total registered cases	2,432,280		1,214,981		1,217,299	

Table 3.A.2 Distribution of confirmed COVID-19 cases, according to disease or comorbidity and sex (Own elaboration)

Federal entity	Totals		Population	Rate per 10^5 inhabitants	Lethality
	Absolute	%			
Aguascalientes	2,428	1.06	1,425,607	170	9.1
Baja California	8,574	3.75	3,769,020	227	17.2
Baja California Sur	1,454	0.64	798,447	182	4.4
Campeche	1,232	0.54	928,363	133	11.9
Coahuila	6,320	2.76	3,146,771	201	9.2
Colima	1,200	0.52	731,391	164	10.0
Chiapas	1,647	0.72	5,543,828	30	14.2
Chihuahua	7,398	3.23	3,741,869	198	13.1
Ciudad De México	34,272	14.98	9,209,944	372	5.2
Durango	2,485	1.09	1,832,650	136	7.2
Estado de México	37,013	16.18	16,992,418	218	14.6
Guanajuato	10,973	4.80	6,166,934	178	8.3
Guerrero	4,490	1.96	3,540,685	127	10.9
Hidalgo	6,238	2.73	3,082,841	202	16.0
Jalisco	12,456	5.45	8,348,151	149	14.2
Michoacán	5,877	2.57	4,748,846	124	12.0
Morelos	3,465	1.51	1,971,520	176	10.2
Nayarit	1,848	0.81	1,235,456	150	15.1
Nuevo León	9,594	4.19	5,784,442	166	7.7
Oaxaca	3,779	1.65	4,132,148	91	8.0

Puebla	11,983	5.24	6,583,278	182	14.0
Querétaro	4,402	1.92	2,368,467	186	6.4
Quintana Roo	2,830	1.24	1,857,985	152	10.1
San Luis Potosí	5,400	2.36	2,822,255	191	8.4
Sinaloa	6,241	2.73	3,026,943	206	15.8
Sonora	6,674	2.92	2,944,840	227	8.8
Tabasco	4,186	1.83	2,402,598	174	6.0
Tamaulipas	5,069	2.22	3,527,735	144	8.4
Tlaxcala	2,503	1.09	1,342,977	186	12.5
Veracruz	9,942	4.35	8,062,579	123	16.0
Yucatán	3,973	1.74	2,320,898	171	9.7
Zacatecas	2,808	1.23	1,622,138	173	9.1
Totals	228,754	100.00	126,014,024	182	9.4

Table 3.A.3 Death rates per 100,000 inhabitants by state (Own elaboration)

Comorbidity	**Total**		**Women**		**Men**	
	Absolute	**%**	**Abso-lute**	**%**	**Absolute**	**%**
Diabetes / Hypertension / Obesity	15,534	6.79	8153	9.50	7381	5.16
Diabetes and hypertension	55,836	24.41	25,044	29.19	30,792	21.54
Diabetes and Obesity	21,322	9.32	10,764	12.55	10,558	7.39
Hypertension and Obesity	27,362	11.96	13,612	15.87	13,750	9.62
Diabetes	84,840	37.09	35,724	41.64	49,116	34.36
Hypertension	102,582	44.84	44,221	51.54	58,361	40.82
Obesity	49,471	21.63	21,978	25.62	27,493	19.23
Total deaths	228,754		85,796		142,958	

Table 3.A.4 Deaths caused by COVID-19, by disease or comorbidity and sex (Own elaboration)

Age group	General		Women		Men	
	Absolute	%	Absolute	%	Absolute	%
0-4	346	0.15	156	0.18	190	0.13
5-14	173	0.08	66	0.08	107	0.07
15-24	1,020	0.45	446	0.52	574	0.40
25-34	5,063	2.21	1,734	2.02	3,329	2.33
35-44	14,683	6.42	4,465	5.20	10,218	7.15
45-54	35,853	15.67	11,934	13.91	23,919	16.73
55-64	58,443	25.55	22,031	25.68	36,412	25.47
65-74	61,403	26.84	23,940	27.90	37,463	26.21
75-84	38,862	16.99	15,592	18.17	23,270	16.28
85->	12,908	5.64	5,432	6.33	7,476	5.23
Total	228,754	100.00	85,796	100.00	142,958	100.00

Table 3.A.5 Deaths caused by COVID-19, by age group and sex (Own elaboration)

CHAPTER IV

EPIDEMIC CURVE OF ESTIMATED CASES OF COVID-19 IN THE STATE OF PUEBLA

MARÍA DE LOURDES SANDOVAL[1]
EUTIQUIO ROMERO[1]
MARCELA RIVERA[1]
LUIS RENÉ MARCIAL[1]
GLADYS LINARES[2]

Abstract

This work presents the epidemic curves of estimated COVID-19 infections in the state of Puebla for different time intervals. Using the estimated weekly infection data published by the Dirección General de Epidemiología de la Secretaría de Salud of Mexico, the data is adjusted using the artificial neural network as a function of cosines. Different adjustments are presented to forecast the behaviour of the epidemic in the following weeks.

Keywords: COVID-19, data fitting, artificial neural networks.

1. Introduction

The COVID-19 epidemic is a social disease. The restraint of contagions depends on the behaviour of each individual in society. It is necessary to reduce the interaction between individuals and to take personal hygiene

[1] Facultad de Ciencias de la Computación, Benemérita Universidad Autónoma de Puebla, 4 Sur 104, Col. Centro, Puebla.
[2] Instituto de Ciencias, Benemérita Universidad Autónoma de Puebla, 4 Sur 104, Col. Centro, Puebla.

measures such as washing hands frequently, wearing a mask and covering a sneeze with the forearm. Also, the reduction of contagions depends on collective actions, such as keeping surfaces disinfected and maintaining the ventilation of closed areas. In order to show the importance of the choice of data to forecast the behaviour of the epidemic, a study of estimated contagions of COVID-19 has been carried out at certain time intervals through data published by the Dirección General de Epidemiología de la Secretaría de Salud of Mexico for the State of Puebla or DGE, for its acronym. To perform the data fitting, an artificial neural network with N weeks was assumed and the epidemic curve was predicted for $N + M$ weeks. $N > 39$, $3 < M < 11$.

The work is structured in sections as described below. In section 2, the neural network fitting technique is presented. In section 3, the graphs with the adjusted data and the predicted data are shown, for different $39 < N < 80$ and $3 < M < 11$. Finally, the conclusions and references used in this work are presented.

2. Fit to an artificial neural network of cosines

Artificial neural networks have been very useful for fitting a data set to a given function. In the next charts, the data of estimated cases of COVID-19 in the State of Puebla on the y-axis are adjusted to a line of series of cosine functions distributed by weeks on the x-axis.

Information on estimated infections is available from week 10 of 2020 to week 35 of 2021, a total of 79 weeks with their corresponding estimated infection cases. The first N data is taken in such a way that x_k represents week k and F_k the number of estimated contagion cases, that is, for $x_1 = 1$, $F_1 = 3$, which corresponds to week 10 of the year 2020; $x_2 = 2$, $F_2 = 35$, corresponds to the information of week 11 of the year 2020 and so on.

k	week/2020	cases	k	week/2021	cases
1	10	3	45	1	4159
2	11	35	46	2	4068
3	12	76	47	3	3487
4	13	128	48	4	2599
5	14	198	49	5	2370
6	15	161	50	6	1897

7	16	282	51	7	1910
8	17	382	52	8	1728
9	18	389	53	9	1755
10	19	564	54	10	1378
11	20	874	55	11	1426
12	21	1123	56	12	1174
13	22	1205	57	13	983
14	23	1686	58	14	1001
15	24	2281	59	15	834
16	25	2932	60	16	650
17	26	2331	61	17	586
18	27	2418	62	18	380
19	28	2591	63	19	375
20	29	2609	64	20	299
21	30	2398	65	21	252
22	31	2014	66	22	244
23	32	1663	67	23	245
24	33	1467	68	24	217
25	34	1426	69	25	322
26	35	1223	70	26	424
27	36	1086	71	27	885
28	37	937	72	28	1270
29	38	921	73	29	2140
30	39	871	74	30	2949
31	40	832	75	31	4095
32	41	899	76	32	4662
33	42	952	77	33	4435
34	43	953	78	34	3627
35	44	951	79	35	3516
36	45	1082			
37	46	1104			

38	47	1263
39	48	1461
40	49	1816
41	50	1946
42	51	2248
43	52	2523
44	53	3284

Table 2-1. Estimated contagion data per week (México s.f.)

With $x = \{x1, x2, x3, \ldots, xN\}$ and $F = \{F1, F2, F3, \ldots, FN\}$ the function can be adjusted. Traditionally, such an adjustment using a neural network is achieved utilizing a supervised multilayer neural network; however, these are inefficient in several cases. In such cases, we can propose the use of a "vectoral" neural network with an activity function that is not an unvarying step function, but a positive defined matrix that allows training, such as the variable learning rate model. (Palma 2015)

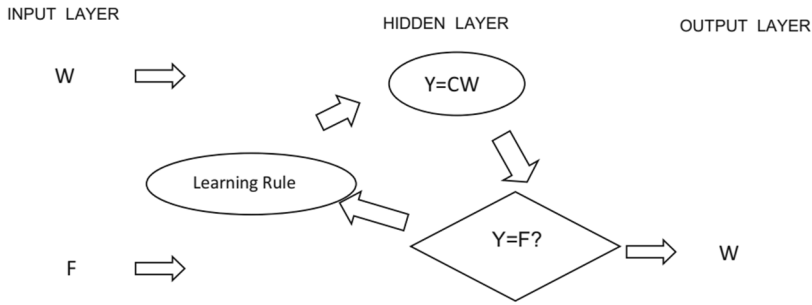


Fig. 2-1. General scheme for the artificial neural network architecture that is used to adjust function Y.

Let $f(x)$ be an unvarying function that we want to adjust and that is known for a finite set of nodes

$$f(x) = W_0 + W_1 cos(x) + W_2 cos(x) + W_3 cos(x) + \ldots + W_N cos(x)$$

and the interpolation criteria must be achieved:

$$f(xk) = W_0 + W_1 cos(xk) + W_2 cos(xk) + W_3 cos(xk) + \ldots + W_N cos(xk) = Fk$$

In the interval $[0, \pi]$

In matrix form

$Y = CW$ with the error $E = F - Y$

Then, the mean square error is:

$$\Phi = {}^{1}/_{2}\,||E'E|| = {}^{1}/_{2}\,||(F-Y)'(F-Y)|| \\ = {}^{1}/_{2}\,||(F-CW)'(F-CW)||$$

To optimize the mean square error, the gradient and Hessian are obtained. (Chaman-Garcia 2011)

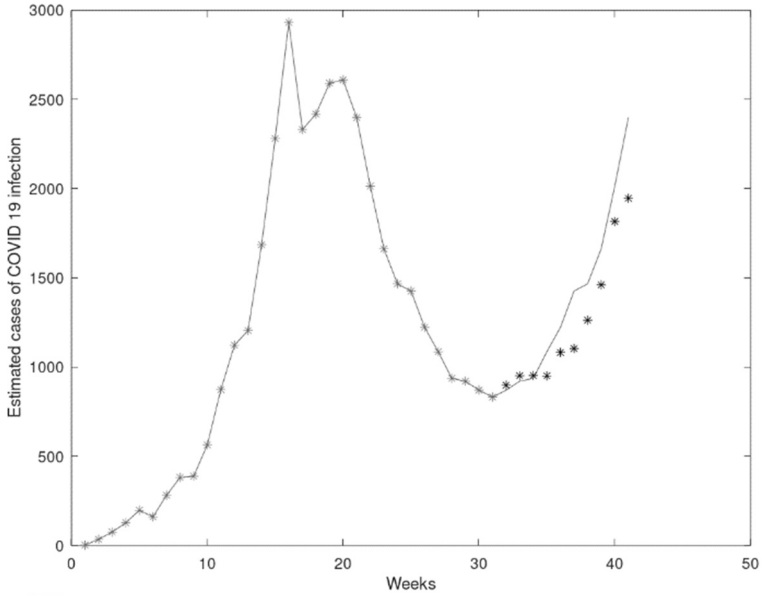
In which the gradient is given by, $g = -CE$ and the Hessian by, $Q = CC'$ Using Newton's method to solve for the minimum of the function Φ, that is: $Qp = -g$,

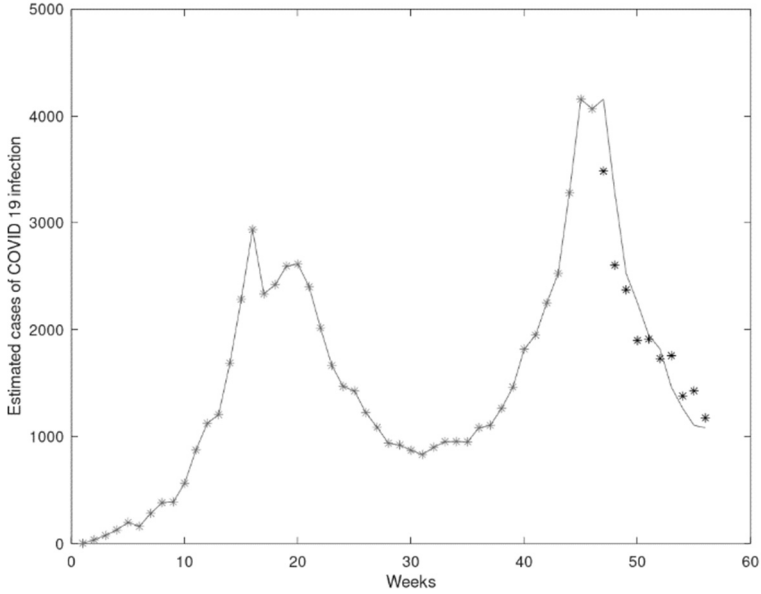
solving this system of linear equations, we have that $W = W_0 + p$, in which W_0 is an initial value given by the user. In the next section, the graphs of the k adjusted data are presented, as well as their prediction up to 10 weeks later, and they are graphically compared with the real data.

3. Application

In this section showing graphs of k, different data adjusted with the cosine neural network are presented, as well as their prediction for 10 weeks afterwards, and it is compared with the real data.

The choice of k is chosen in the ascending curve and in the descending curve to know under which conditions good predictions can be had. To test the performance of the proposed interpolation, an adjustment of the information of the first k weeks of estimated infections of COVID-19 infections in the state of Puebla is considered. Subsequently, the next 10 weeks are predicted and compared graphically with the real data. Two behaviours are observed: with good prediction, four graphs corresponding to the start of the pandemic until the weeks 31, 46, 65 and 76 respectively.


Estimated cases of COVID 19 infection
Weeks
0
500
1000
1500
2000
2500
3000
0
10
20
30
40
50


Estimated cases of COVID 19 infection
Weeks
0
1000
2000
3000
4000
5000
0
10
20
30
40
50
60

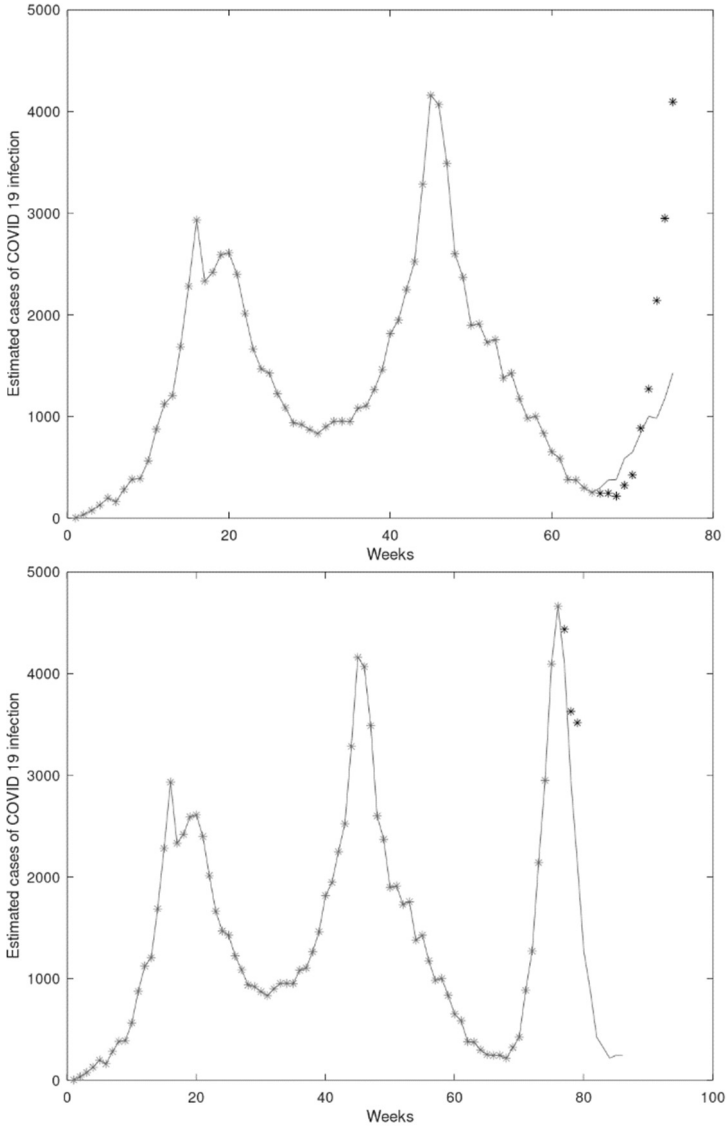
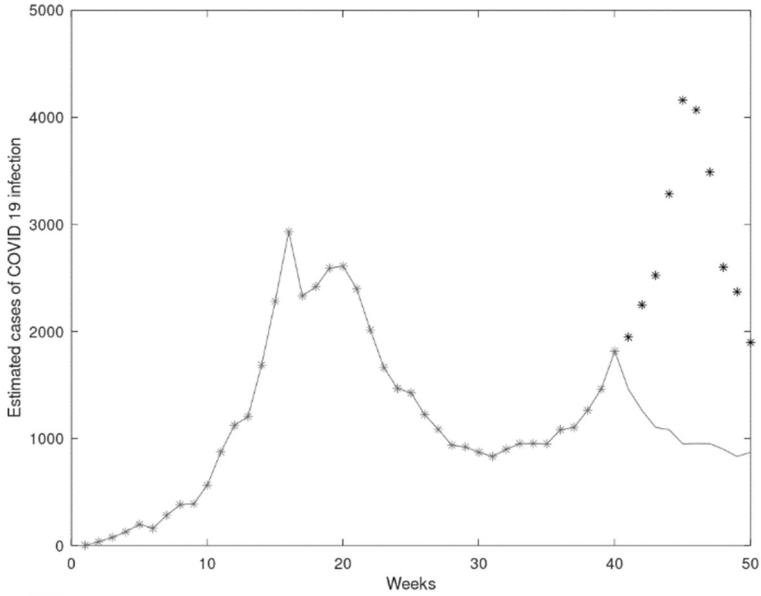
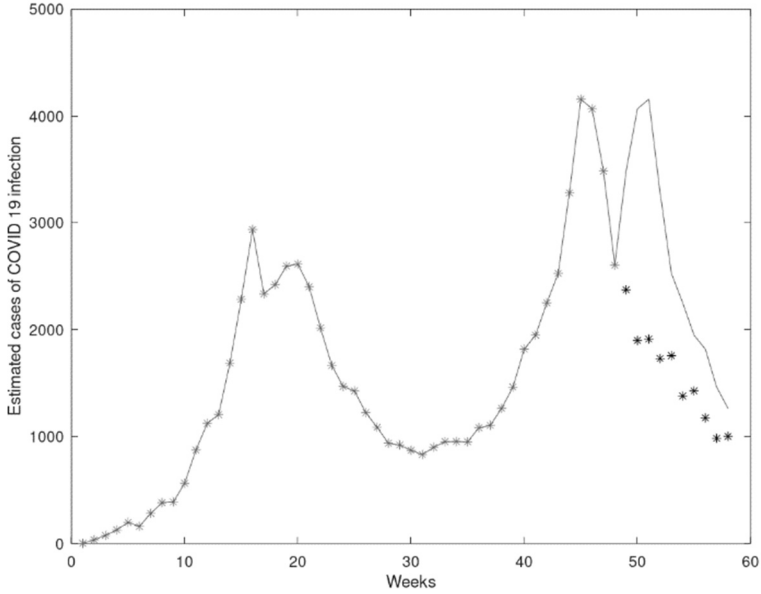
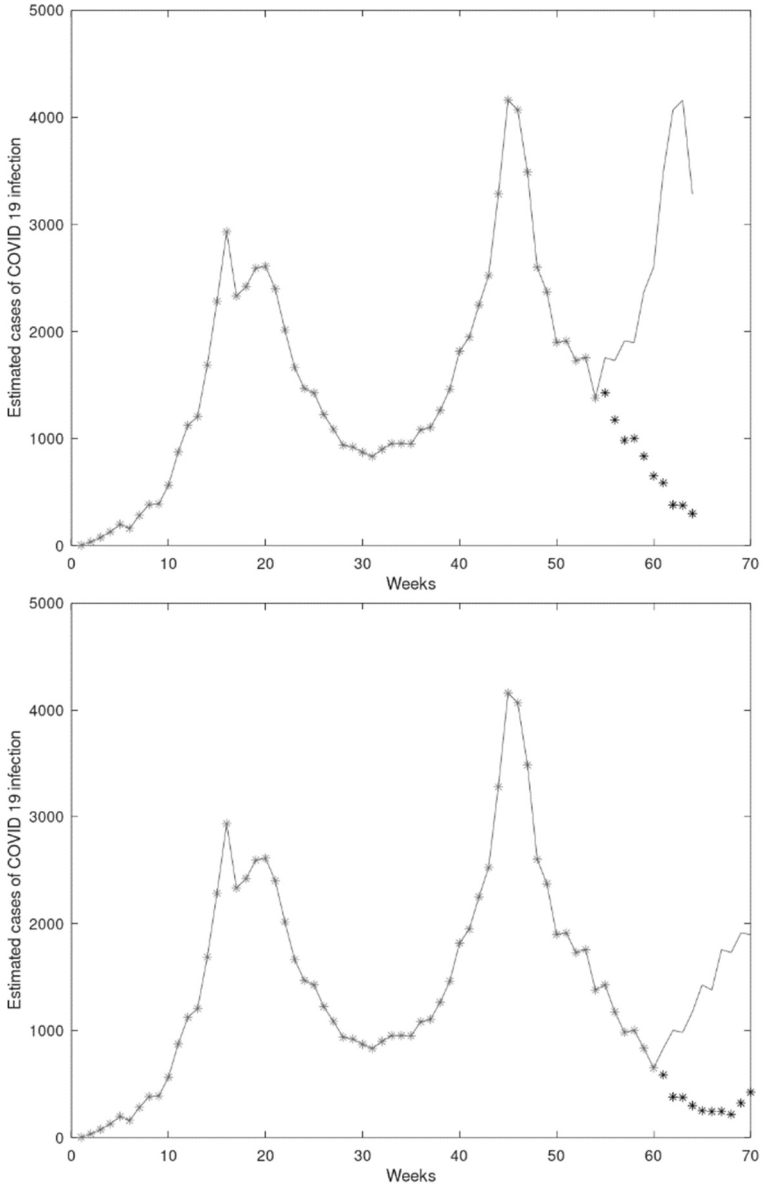


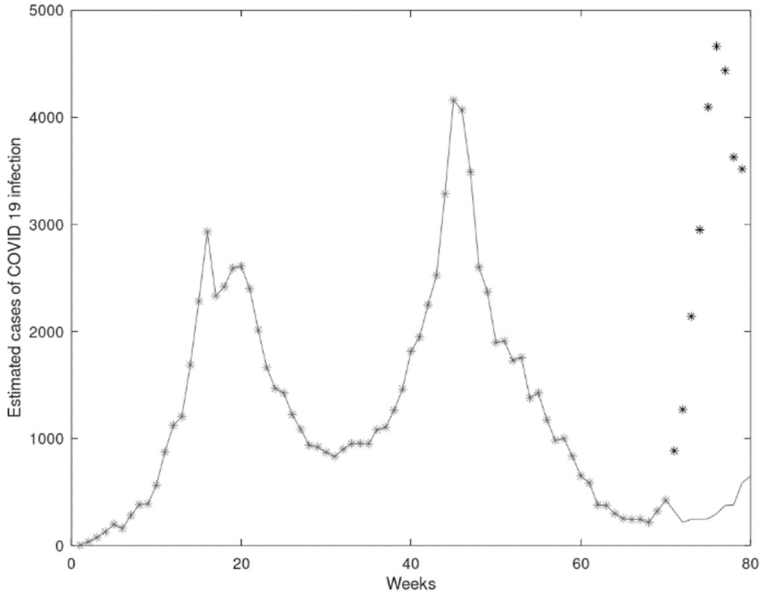
Fig. 3-1. From top to bottom, the graphs corresponding to weeks 31, 46, 65 and 76 of the good predictions of estimated infection cases by COVID-19.

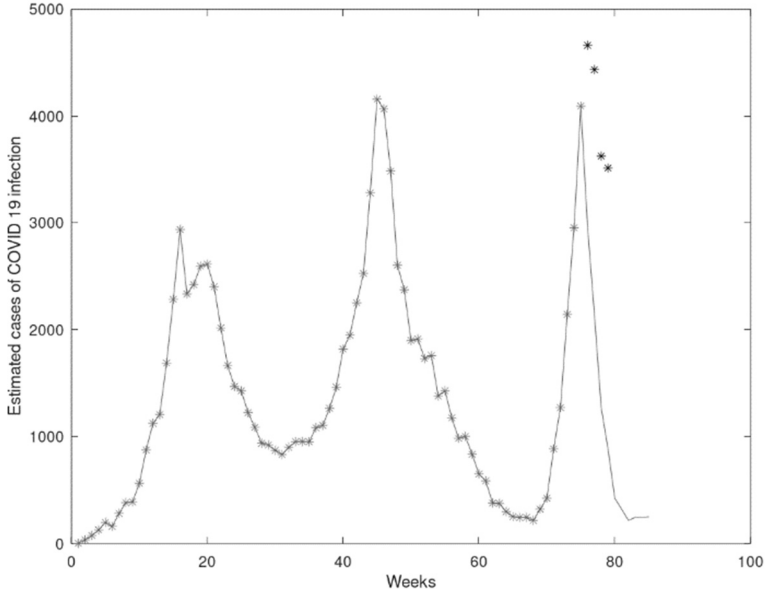
With bad prediction, there are seven graphs corresponding to the start of the pandemic until the weeks 40, 48, 54 60, 70, 75 and 79, respectively.


Estimated cases of COVID 19 infection
Weeks
0
1000
2000
3000
4000
5000
0
10
20
30
40
50


Estimated cases of COVID 19 infection
Weeks
0
1000
2000
3000
4000
5000
0
10
20
30
40
50
60


Estimated cases of COVID 19 infection
Weeks
0
1000
2000
3000
4000
5000
0
10
20
30
40
50
60
70


Estimated cases of COVID 19 infection
Weeks
0
1000
2000
3000
4000
5000
0
20
40
60
80


Estimated cases of COVID 19 infection
Weeks
0
1000
2000
3000
4000
5000
0
20
40
60
80
100

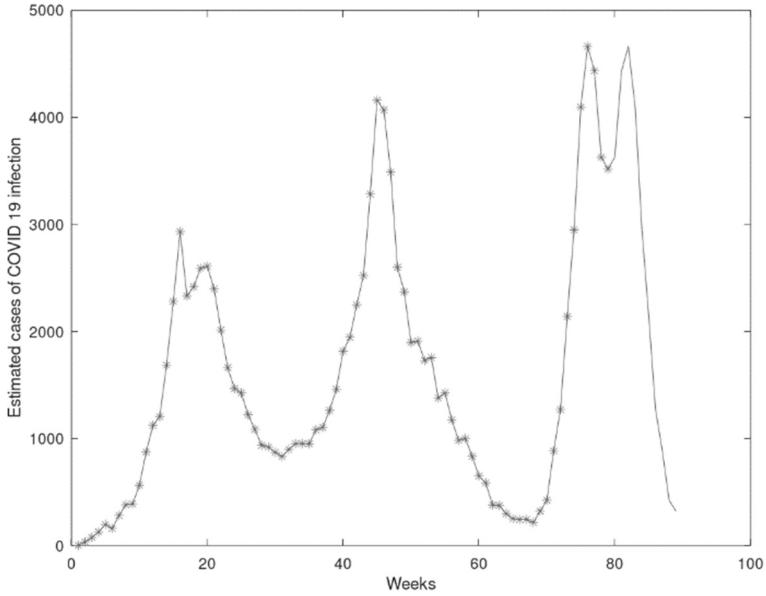


Fig. 3-2. From top to bottom, the graphs corresponding to weeks 40, 48, 54 60, 70, 75 and 79 of the bad predictions of estimated infection cases by COVID-19.

Conclusions

The interpolation was made using the contagion data estimated by COVID-19 of the first k weeks. The interpolation was used to predict the behaviour of the epidemic ten weeks later. It is observed that a good prediction is obtained when the interpolation is carried out in weeks $1, 2, \ldots, k$; in which k corresponds to contagions estimated maximum or minimum of the epidemic curve.

Bibliography

Chaman-Garcia, Ivan, and Lourdes Sandoval-Solis. 2011. «Ajuste e Integracion Numerica de Funciones Muy Oscilantes usando Redes Neurona Les Artificiales supervisadas.» *Investigación Operacional* 193.

México, COVID 19. s.f. *COVID 19 Tablero México - Dirección General de Epidemiología.* Último acceso: 20 de 09 de 2021. https://coronavirus.gob.mx/datos.

Palma, L. Sandoval and José M. Amero and Gabriel J. Vázquez and Alejandro. 2015. «Franck–Condon factors using supervised artificial

neural networks. I. The CF+ cation.» *Journal of Molecular Modeling* 1-4.

CHAPTER V

ON SAMPLE SELECTION IN NETWORKS AND THE MODELLING OF EPIDEMIC ISSUES

CARLOS BOUZA[1]
SIRA ALLENDE[1]
RICARDO KALID[2]
RILTON PRIMO[2]

1. Introduction

When modelling epidemics, it is possible to assume the existence of a certain compartmenting, of individuals interacting homogeneously within groups. At first, a small number of people is infected, and the spread of the virus may be modelled considering some unknown probability model. Contact patterns among individuals generate the spread of the sickness. The epidemiologists are interested in modelling these patterns. Diekmann-Heesterbeek (2000) discussed the role of probability models in this context. For example, in the COVID-19 pandemic, as in many others, contamination is spread by person-to-person contacts. The first infected individuals act as seeds of a chain of contacts with other people. Some of them are infected and an unknown probability measure generates the rate of transmission.
Assuming that there is a homogeneous mixing of infected and susceptible healthy people is not a correct pattern, generally. The evidence of the epidemics of SARS, 2002–2003, provided evidence of the righteousness of this statement, see Ferguson (2005), Kucharski et al. (2020), Longini (2005), Zhigljavsky et al. (2020), Tang et al. (2020). They assumed the reproductive value of this epidemic was smaller than supposed previously.

[1] Universidad de La Habana, Cuba
[2] Universidade Federal Do Sul Bahia, Brazil

When the pattern is heterogeneous, the transmission process may be modelled as a network. The graph is described considering individuals as nodes and contacts as edges.

Graphs Theory is used in different contexts. Nowadays the main areas of applications are describing computers, communication and social networks, see Tang, et al. (2020), Lloyd (2001), Ahn, et al. (2007), Chin, et. al. (2020) and Wilson et al. (2009). In epidemiology, graphs are becoming a useful tool (see Meyers et al. (2005) and Newman (2002)). This theory allows the evaluation of different aspects of the dynamic of an epidemic. Individuals are nodes and contacts are edges, which are activated when a contact between two nodes is detected. Contamination from an infected to a healthy person takes place with a certain probability. The number of infected people and other parameters of the epidemic, under set of conditions, determines graph structures that may be of use for evaluating, predicting and modelling the number of expected contaminated persons, the risks of being in a contagious chain, etc.

Following the health status of recovered patients is of importance for establishing the frequency of certain sequels, for example. Many retrospective studies on clinical characteristics of people recovered from SARS-CV-2 have been developed. For example, they look for regularities on parameters of Ct, anti-SARS-CoV-2 IgG, antibodies IgM in COVID-19's recovered patients, the difference of anti-SARS-CoV-2 IgG and antibodies IgM, etc – see An et al. (2020) for a technical discussion. Using the graph structure of the populations for sampling would provide an adequate statistical frame for post-pandemic studies.

Coronavirus (COVID-19) is spread mostly based on social contacts in groups. Epidemiologists must deal with studies for establishing regularities on the potential widespread of the contagion. Strategically, they must cope with the existence of uncertainty. They should use statistical tools for collecting and processing data for and informing and evaluating the effect of policies in its mitigation. Common tools available are not sufficiently accurate and robust modelling is needed for efficient policy making. Flexible modelling should be considered, for increasing accuracy. Stochastic models used should capture the probabilistic nature of contagion between the member of a population (e.g., communities, regions, etc.). The models should assess parameter sensitivity for the variables that characterise the COVID-19 pandemic. The number of infected persons at a fixed time is the main quantity of interest in pandemic studies. In post-

pandemic studies, the sequels of the treatments, aspects of the recovery etc. are the interest quantities.

Graphs are made up by a series of characteristics as degree distribution (see Ribeiro and Towsley (2012), Stumpf and Wiuf (2005)). Some properties are more important for establishing key aspects of the epidemic graphs. Understanding the network provides insights on the behaviour of an epidemic. Having characterised the graph status, researchers may further develop criteria for controlling the spread of it in pandemic studies. Post-pandemic, epidemiologists may use the model for evaluating the effect of treatments (sequels, correlations with demographic variables, etc.), so important health problems may be described in their dynamic interrelations. Of course, the generalizations depend on whether the properties are preserved when transforming a graph or not – see Moitra (2009) for discussing of preservation, running an algorithm on the transformed graph allows and on considering the behaviour of the original graph. That is, the properties of an original graph G may be studied by analysing the transformed graph G_s.

In real life, a simple and effective method is studying the original graph by using as a transformation method some sampling procedures. Sampling a graph provides a fast and cheap way of transformation. Sampling methods are very popular in applications, due to their efficiency (see the basic elements in Lavallèe (2007), Fuller (2009), Arnab (2017)). The sample graph size is smaller and usually exhibits similar properties to the original one. The sample graph is expected to be representative. Graph sampling procedures are simple and intuitive. It is expected that the first thoughts of a researcher are to handle massive data by obtaining a sample. As the epidemiologist investigates people and their relation, the aims of their investigation deals with selecting a sample of vertices or edges from G.

The clients consulting the statistician are looking to obtain advice on the sample size, how to select the sample, how to elaborate inferences and how to measure the efficiency of the proposed procedures. The statistical framework considers the need to construct a population frame $U = \{u_1, \dots, u_N\}$ for the particular problem. In some occasions there are contacts among the members of U, say there are certain connections, and a graph may represent the population. Taking

$$I(u_i, u_j) = \begin{cases} 1 \; if \; u_i \; is \; connected \; with \; u_j \\ 0 \; otherwise \end{cases}$$

a graph G is modelled: individuals are nodes (vertices) and the relations are edges.

This framework may be assumed in many common applications in communication studies, social networks, etc. In medicine the needs for modelling a virus propagation, the characterization of immunization policies, etc, are similar. That is present in strategic studies of the spread of the COVID-19 virus. An epidemiologist may ask for sampling methods, where the network structure is taken into account for obtaining information on relations among individuals, in studies of the transmission, estimating risks, detecting important sources of transmission as well as other epidemiologic issues.

The epidemiologists are more interested in obtaining data for a population graph using simple techniques. Sampling in epidemiology has been widely used but the performance of graph sampling has not been systematically studied.

2. Some ideas on Epidemiological Graphs

Epidemiological studies suffer from the lack of data. Using a sample of people provides an approach that allows the epidemiologist affordance to draw people from a population of possible infected individuals. Picking from identified people in a Health Service, for example in a municipal, may be implemented by considering that the people are vertices of a network. The sample should permit to answer questions as how many vertices (people) or edges (connections between people) should be sampled in order to obtain an appropriate coverage. Note that is surveying hidden populations. The epidemiologist commonly wants to reach the hidden population of infected people but without symptoms. It is impossible to enumerate and sample all the people. A recommendation is to start from a small set of individuals considering them as seed nodes. The set of individuals to be surveyed is expanding according to the knowledge obtained in the process. Salganik and Heckathorn (2004) and Yan and Gregory (2013) present elements needed for understanding the problems of dealing with hidden populations.

The populations under study constitute complex and large networks. The corresponding graphs are hard to manipulate using the whole graph. This poses the existence of graph sparsification, a classical problem in terms of edge, vertex or both (see discussions in Harvey (2011), Moitra (2009)). The

determination of interactions between all the individuals in a neighbourhood. Generally, it is too costly to sample the graph and perform PSR tests on infected individuals as well as to determine the active edges on the sampled graph.

In epidemiological studies the type of network is not specified as in other real-life networks. Such researches, commonly, lack knowledge of which is the underlying graph. The investigator may assume that certain type of graph is an acceptable model.

Nowadays, in real life applications, the populations under study constitute complex and large networks (see the basics on networks in Newman (2010)). The corresponding population graphs are hard to manipulate. This poses the existence of graph sparsification, a classical problem, in terms of edge, vertex or both. The determination of interactions between all the individuals in a neighbourhood generally is too costly to sample the graph and perform PSR tests on infected people and determine the active edges on the sampled graph. Commonly the epidemiological graph is too large to fit in a screen and sampling permits to obtain an easier visualization. The epidemiologist must deal with some mathematical issues as:

- Effect of graph size reduction during the transformation.
- Expected representativeness, similarity to the population graph, etc.

The common issue is to reduce the graph while size preserving its properties. Theoretically, a mathematical programming problem is posed for minimizing the distance between populations and sampled graphs, see Birnbaum et al. (2008). This problem is extremely costly in epidemiological studies. In practice, running the sparser may be more complex than running an algorithm on the original graph. From this quotation, the epidemiologist will be more interested in using a simple technique and evaluating the results in the sampled graph. For them, a graph sampling procedure should be based on simple and intuitive approaches. With these observations, we are more interested in simple techniques to see what results can be obtained. Graph sampling is a simple and intuitive approach. It is usually one of the first thoughts to handle massive data. It has been widely used, but the performance is not systematically studied.

The common issue in real life research is to reduce the graph size preserving its properties. Theoretically a mathematical programming problem is posed for minimizing the distance between population and sampled graphs. This

approach is extremely costly in epidemiological studies. Note that reducing the graph size preserving the cuts is NP-Hard. Polynomial approximation algorithms are too complex.

In practice, running the sparsifier is more complex than running an algorithm on the original graph. The researcher expects that the sampled graph "looks like" the population graph and the satisfaction of the Dimensionality Reduction and Graph Embedding is also considered. From these quotations, the epidemiologists will be more interested in using a simple technique and evaluating the results in the sampled graph. For them, a graph sampling procedure should be based on simple and intuitive approaches. With these observations, we are more interested in simple techniques to see what result can be obtained. Graph sampling is a simple and intuitive approach. It is usually one of the first thoughts to handle massive data. It has been widely used, but the performance is not systematically studied.

3. Some needed basic ideas on sample graphs:

The best approach for obtaining a sample graph is sampling to obtain a representative subset of vertices. For example, it seems this approach is used when identifying COVID-19 asymptomatic individuals. Note that the sampled vertices are people. In epidemics, the target population usually is hidden. Then, the researchers select some sampling algorithm and apply it to identify the hidden population. For example, a sample is selected and people in it are tested for COVID-19 by using a PSR test. See Ferguson et al. (2020), Pham, (2020) for examples on potential application areas. Then, the hidden population of individuals carrying the virus has been subsampled.

The objective in sampling graphs is to estimate properties of *G, see* Wang et al. (2011). Hence, the researcher expects that the property is preserved on G_s. Say, given the unknown epidemiological graph G, the researcher defines a property. It is a function $f(G)$, probably a vector function. Once observed G_s is calculated $f(G_s)$. Clearly $f(G_s)$ is random if G_s is generated randomly. It is an estimator of $f(G)$.

Motivations of property estimation and property preservations apparently are different. They really are closely related and can be sometimes transformed to each other. When the goal of the investigation is only

estimation, the preservation of the properties on G_s are less important. Knowing that the properties are biased, a statistical correction is possible.

Some basic issues on Graph Theory should be revisited. Some of them are presented below.

- The sampled graph by $G_s = (V_s , E_s)$, satisfies the conditions
$$V_s \in V$$
$$E_s \in E$$

- These conditions ensure that vertices or edges are a sample from original graph
$$E_s \subseteq \{v \cdot v | u \in V_s; u = v \cdot v\}$$

- These conditions ensure that the sampled elements determine a valid graph.

- Given *G,* the researcher defines a property. It is a function *f(G),* probably a vector function. Clearly $f(G_s)$ is random.

From a statistical point of view, satisfying some property preservation supports that the inference procedures may be optimal. Unfortunately, in the presence of properties of estimators do not preserve necessarily property results. Nevertheless, in estimation, direct property without deriving provisory property preservations result simplifies the analysis of the behaviour of inferential procedures.

A graph may be characterised using different sets of properties. Graphs sharing the same properties are considered "similar". The researcher may rely on the similarity among graphs for observing or estimating common properties. Using this fact, the epidemiologist analyses a simpler graph than the population one. For example, something may be said about the behaviour of the spread of the COVID-19 virus or on particularities of the evolution of recorded patients in post-pandemic studies (Müller et al., 2000).

4. Graph Structure and its use in COVID19 problems

Some patterns of the original, and generally unknown, graph G allow the identification of some parameters of epidemiological interest for

characterising the behaviour of the COVID-19 pandemic in a particular population. Among them are:

Take some general characteristics of graphs that are of use in epidemiology studies.

- Network Size. It is described by
  The number of vertices (individuals): $N = |V|$
  The number of edges (connections): $M = |E|$.
  The degree of the node *v: d(v).*

Take $w(v) \in R$, $v \in V$ as a weighting function; $\delta(S) = \{(u, v) \in E | u \in S, v \notin S\}$; $vol(S) = \bigcup_{v \in S} \delta(v)$.

Note that in epidemiological studies the number of individuals in the network (*N*) are of interest, the number of connections among them (*M*) and how many people are connected with other *v (d(v)).*

- The cut of a set S is the edge crossing S and V − S:
$$Cut(S) = |\delta(S)| = |\{(u, v) \in E | u \in S, v \notin S\}|$$

It permits the analysis of how many individuals with the virus, detected in a group S, relate to people working in another out of the group.

When the cuts are preserved on G_s, weights of the cut may be evaluated and used as a naive and natural estimator for it in *G*. In the opposite case, cuts are not preserved on G_s. Then G_s should be transformed so that the cuts are preserved. Consider that the interest is on the connections among individuals and the edges are sampled with a fixed probability *p*. It is clear that the cuts of G_s have less edges than *G*. Up-weighting all the edges in G_s, by a factor of p^{-1}, is enough for preserving the weighted.

Related measures are:

- The ratio of S-cuts $RCut(S) = \frac{|\delta(S)|}{|S|}$ ($w(v) = 1, \forall v \in V$). It is the relative importance of the event detected in S for the rest of the workers of the office.

- The normalized cut of S: $Ncut(S) = \frac{|\delta(S)|}{|vol(S)|}$ ($w(v) = d(v), \forall v \in V$). It is the relative importance of the event detected in S for the importance of S.

- The weighted cut of S: $Cut(S) = \frac{|\delta(S)|}{\sum_{v \in S} w(v)}$. It is the relative importance of the event detected in S for the rest of the workers of the office in terms of the assigned weights.

Consider the original graph G and the sampled graph G_s and use cut weights. When the cuts are preserved on G_s, the weights of the cut may be evaluated and used as a naive and natural estimator for it in G. In the opposite case, cuts are not preserved on G_s, a transformation should be needed, so that they are preserved. For example, if the interest is on the connections among individuals and the edges are sampled with fixed probability p, naturally the cuts of G_s have less edges than G. In up-weighting all the edges in G_s by a factor of p^{-1}, the weighted cuts are preserved.

As previously quoted, the density of the degrees is of importance. The epidemiologist would be very interested in studying the proportions of the number of people involved with other ones. This is of particular importance when a person carrying the virus k contacts density. Based measures of importance are:

- Degree Distribution (Deg). Draw randomly a node $u \in V$, the p.d.f. for the degree distribution is

$$P_{Deg}(k) = Prob(d(u) = k)$$

The use of the sample proportion

$$p_{Deg}(k) = \frac{1}{|V_s|} \sum_{u \in V_s} I(d(u) = k)\,;\; I(d(u) = k) = \begin{cases} 1 \; if \; I(d(u) = k) \\ 0 \; otherwise \end{cases}$$

provides unbiased estimations with error $\frac{P_{Deg}(k)(1-P_{Deg}(k))}{|V_s|}$.

- Power Law Exponent (PLE). Once derived $p_{Deg}(k)$, γ

- Power law exponent =$p_{fit}(k) \propto k^{-\gamma}$.

It is the closest to the observed $p_{Deg}(k)$.

- Graph Density (GD) is the ratio of the observed number of edges over the maximum possible number of them:

$$GD = \frac{M}{\binom{N}{2}} = NE\big(d(X)\big) = N\sum_k kP_{Deg}(k)$$

Consider $\widehat{GD} = N\sum_k kp_{Deg}(k)$. $p_{Deg}(k)$ as the involved variables are binomials, as it is a good estimator of $P_{Deg}(k)$. The asymptotic normality of the estimator follows from the properties of the binomial distribution. $\widehat{GD}$ is a linear function of asymptotical normal distributed random variables and its error $V(\widehat{GD}) = N^2\sum_k k^2\frac{P_{Deg}(k)\left(1-P_{Deg}(k)\right)}{|V_S|}$ is estimated by $\hat{V}(\widehat{GD}) = N^2\sum_k k^2\frac{p_{Deg}(k)(1-p_{Deg}(k))}{|V_S|}$.

Therefore, the T-Student statistic to be used in inferences is

$$T(GD) = \frac{\widehat{GD} - GD}{\sqrt{N^2\sum_k k^2\frac{p_{Deg}(k)(1-p_{Deg}(k))}{|V_S|}}}$$

Their potential applications in COVID-19 are evident. The epidemiologists are interested in evaluating the proportion of individuals with k connections. For example, considering the contamination, it is important to establish the proportion of individuals with more than h contacts. That is $\sum_{k>h} p_{Deg}(k)$, the expected number of contagious $E(\,d(X)) = \sum_k kp_{Deg}(k)$. If a theoretical distribution is identified, the inferences are going to be more accurate, frequently. *PLE* provides a good fit for the theoretical distribution in many occasions in epidemic research. Evaluating, *GD* provides the worst scenario.

Take

- $\sigma_{u,i}(v) = \{p \in \sigma_{u,i} | v \in p\} = paths\ where\ v\ is\ involved$

- $\sigma_{u,i} = collection\ of\ shortest\ paths\ between\ u\ and\ i$,
- $q_k = \frac{(k+1)p_{\mathrm{Deg}}\ \ (k+1)}{\sum_j\ jp_j}$ (distribution of edges excepting the one linking the two considered vertices),
- $p_{j,k}$ = joint distribution of the degree of the vertices j, k;
- $\Delta(v) = \{(u, w)|u \in N(v), w \in N(v), (u, w) \in E\}$ (set of observed edges in the neighbourhood .
- $N(v)$); $s(G) = \sum_{u,v \in V} d(v)d(u)$.

With these definitions, the behaviour of the movements from a node to another are based on elating paths. In epidemics as such COVID-19, the movements denote the spread of the virus. Useful path measures are:

- Path Matrix (PM). $[P_t]_{i,j} = \begin{cases} 1\ if\ \exists [A^\tau]_{i,j} > 0 \\ 0\ otherwise \end{cases}$. A is the adjacency matrix, the path matrix encodes the information reachability between any pair of vertices.
- Shortest Path Matrix (SPM)= $[S]_{i,j} = arg\ min_t [P_t]_{i,j}$(the shortest path matrix). It records all shortest path distance of pairs. The collection of shortest paths between *i* and *t* is denoted by $\sigma_{i,t}$.
- Average Path Length (APL)=$\frac{1}{\binom{N}{2}} \sum_{1 \le i < j \le N} \quad arg\ min_t [P_t]_{i,j}$.
- Radius (R) for a fixed $v = R(v) = Max_{u \in V}\ S_{u,v}$. It is the maximum shortest path distance from *v* to all the other nodes in V. A small value indicates that, if the individual represented by the node *v* is infected, the spread with smaller paths is no larger than *R(v)*.
- Diameter (Dia) for a fixed $v = = DG(v) = Max_{v \in V}\ R_v$. It is the longest distance between all pairs of vertices. Generally, it is more useful using the effective diameter (the distance within which α fraction of the pairs can connect with each other). An interpretation goes in the same line as with *R(v)*.

These measures are defined in terms of an extreme value. The epidemiologists may use these measures in a graph sampled for evaluating the spread of the virus. Particularly, naive estimators are:

$$\widehat{APL} = \frac{1}{\binom{|V_s|}{2}} \sum_{1 \le i < j \le |V_s|} arg\ min_t [P_t]_{i,j}$$

$$\hat{R}(v) = Max_{u \in V_s}\, S_{u,v}$$

$$\widehat{DG}(v) = Max_{v \in V_s}\, R_v$$

The first estimator is a sample mean and the CLT may be used for deriving inferences, but it is biased for *APL.* A consistent estimator of the MSE is

$\varepsilon(\widehat{APL}) = \frac{1}{\binom{|V_s|}{2}} \sum_{1 \le i < j \le |V_s|} (arg\ min_t [P_t]_{i,j} - \widehat{APL})^2$. Using these facts may be considered the asymptotic normality of $T_{APL} = \sqrt{\frac{\widehat{APL} - APL}{\varepsilon(\widehat{APL})}}$

In the other cases naive estimators of the MSE are

$$\varepsilon\left(\hat{R}(v)\right) = \frac{1}{|V_s|} \sum_{u \in V_s} (Max_{u \in V_s}\, S_{u,v} - \hat{R}(v))^2$$

$$\varepsilon\left(\widehat{DG}(v)\right) = \frac{1}{|V_s|} \sum_{u \in V_s} (Max_{v \in V_s}\, R_v - \widehat{DG}(v))^2$$

For evaluating the behaviour of the estimations of the ratios and diameters, the estimated MSEs may be averaged. That is, the epidemiologists select a sample graph and compute it

$$\bar{\varepsilon}(PM) = \frac{1}{|V_s|} \sum_{v \in V_s} \varepsilon(PM(v));\ \varepsilon(PM(v)) = \varepsilon\left(\widehat{DG}(v)\right), \varepsilon\left(\hat{R}(v)\right)$$

Therefore, an analysis of the distribution of the degrees of G may be performed computing the sample means for $PM = R, DG$:

$\overline{\widehat{PM}}=\frac{1}{|V_s|}\sum_{v\in V_s}\quad \widehat{PM}(v)$ and $\bar{\varepsilon}(PM)=\frac{1}{|V_s|}\sum_{v\in V_s}\quad \varepsilon\left(\widehat{PM}(v)\right)$.

A question is if

$$T_{PM\varepsilon}=\frac{\widehat{PM}-\overline{\widehat{PM}}}{\sqrt{\varepsilon(\widehat{PM})}};PM=R,DG$$

provide good speed of convergence to the N(0,1).

The epidemiologists may be interested in evaluating $\hat{R}(v)\ and\ \widehat{DG}(v)$ for certain individual v but in the studies on the dynamic of COVID-19, the mean of the paths measures are of main interest. That is to evaluate the values of

$$\bar{R}=\frac{1}{|V|}\sum_{v\in V}\quad R(v)=\frac{1}{|V|}\sum_{v\in V}\quad Max_{u\in V}\,S_{u,v}$$

$$\overline{DG}=\frac{1}{|V|}\sum_{v\in V}\quad DG(v)=\frac{1}{|V|}\sum_{v\in V}\quad Max_{v\in V}\,R_v$$

The sample means perform as good estimators. Then may be used

$$\hat{\bar{R}}=\frac{1}{|V_s|}\sum_{v\in V_s}\quad R(v)$$

$$\widehat{\overline{DG}}=\frac{1}{|V_s|}\sum_{v\in V_s}\quad DG(v)=\frac{1}{|V_s|}\sum_{v\in V_s}\quad\sum_{v\in V}\quad Max_{v\in V}\,R_v$$

Both estimators may be considered as convergent to a normal distribution under large value of $|V_s|$ and the inferences will be based on the T-Student statistic

$$T_{PM}=\frac{\widehat{PM}-\overline{\widehat{PM}}}{s(\widehat{PM})};PM=R,DG,s\left(\widehat{PM}\right)$$

$$=\sqrt{\frac{1}{|V_s|(|V_s|-1)}\sum_{v\in V_s}\quad\left(R(v)-\hat{\bar{R}}\right)^2}$$

5. Measures of similarity

The existence of similarities may be evaluated by means of ideas coming from statistical taxonomy and clustering. Some of them are:

- Closeness Centrality (CloC)=$Closeness\ (v) = \frac{1}{\sum_{u \in V} S_{u,v}}$.
- Between-ness Centrality (BC)= Between-ness (v)= $\sum_{u \neq v, i \neq v} \frac{|\sigma_{u,i}(v)|}{|\sigma_{u,i}|}$. It is the number of shortest paths where the person *v* is involved in. A larger value of *BC(v)* with respect to *BC(u)* indicates that the dissemination of the virus by v is faster than the contamination due to *u*.
- Assortative (As)= $r = \frac{\sum_{j,k}(p_{j,k} - q_j q_k)}{\sigma_q^2}$. It is the Pearson Correlation of similarity between neighbouring vertices.

These measures, when evaluated in a graph sample, are:

$$CloC\ (v|G_s) = \frac{1}{\sum_{u \in V_s} S_{u,v}}$$

$$BC(v)|G_s = \sum\nolimits_{u \neq v, i \neq v | G_s} \frac{|\sigma_{u,i}(v)|}{|\sigma_{u,i}|}$$

$$r(G_s) = \frac{\sum_{j,k}(p_{j,k} - q_j q_k)}{\sigma_q^2}$$

- Clustering Coefficient (CC)=$C(v) = \frac{|\Delta(v)|}{\binom{|N(v)|}{2}}$. $\Delta(v)$, $\binom{|N(v)|}{2}$

Note that $|\Delta(v)|$ is the number of triads determining triangles and *N(v) is* the number of possible edges in the neighbourhood. Hence, C*(v)* is some type of rate, say the ratio of the number of observed triangles (connections among 3 individuals) divided by the number of possible triangles on the *v*'s in the *v*'s network.

They give important information on the similarity assigned to the person v in the graph. See a discussion in Friggeri et al. (2011). The similarity is characterized by the means

$$\overline{MS} = \frac{1}{|G|}\sum_{v\in G} MS\,(v|G); MS = CloC, BC, r, CC$$

which are estimated using the naive estimators

$$\widehat{\overline{MS}} = \frac{1}{|G_s|}\sum_{v\in G_s} MS\,(v|G_s); MS = CloC, BC, r, CC$$

The measure of the error

$$\varepsilon\left(\widehat{\overline{MS}}\right) = \frac{1}{(|G|-1)|G_s|}\sum_{v\in G}(MS\,(v|G) - \overline{MS})^2$$

is estimated naively by

$$\hat{\varepsilon}\left(\widehat{\overline{MS}}\right) = \frac{1}{(|G_s|-1)|G_s|}\sum_{v\in G}(MS\,(v|G) - \widehat{\overline{MS}})^2$$

As $\widehat{\overline{MS}}$ is a mean, the researcher may consider that normal approximation is laid and use for inferences the statistic

$$T_{MS} = \frac{\widehat{\overline{MS}} - \overline{MS}}{\sqrt{\hat{\varepsilon}\left(\widehat{\overline{MS}}\right)}}$$

- Average Clustering Coefficient or Global Clustering Coefficient. It summarizes the CC of all vertices of G:

$$C(G) = \frac{1}{|V|}\sum_{v\in V} C(v)$$

As it is a mean and an unbiased estimator and its error are

$$\hat{C}(G_s) = \frac{1}{|V_s|}\sum_{v} C(v \in V_s)\,; V(\hat{C}(G_s))$$

$$= \frac{1}{|V_s|(|V|-1)}\sum_{v\in V}(C(v) - C(G))^2$$

the CLT seems to be acceptable, hence for inferences is considered as a good method to consider that.

$T_G = \frac{\hat{C}(G_s)-C(G)}{\sqrt{\widehat{V}(\hat{C}(G_s))}} \sim N(0,1)$; $\widehat{V}(\hat{C}(G_s)) = \frac{1}{|V_s|(|V_s|-1)}\sum_{v\in V_s}\left(C(v) - \hat{C}(G_s)\right)^2$.

6. Sampling designs

6.1. General issues

The objective generalization of research commonly relies on proving statistically propositions about theoretical hypothesis on the issues of interest. For performing a proof one needs to know a probability distribution or elaborate data-based inferences. See discussions in Sarlós (2016). Statistical models allow extracting from the data ideas on the involved errors, lack of fit of theoretical structures and supporting elaborated theories derived from the study. Adoption of a probabilistic approach permits to determine whether sample graph supports that the network adheres to previous epidemic conceptions.

The questioning of epidemiologists, at individual level, may be studied through the degrees, in-degrees, out-degrees, mutual relations and closeness (clustering). At a local level the dyads and triads may be used for considering equivalences, transitivity and clustering. At a global level the network's properties supporting the analysis include the study graph sampling. It has received a lot of attention recently due to the needs of studying networks in computation, internet, economics, social media etc.

Classical approaches are random sampling of vertices or edges, random walks, and random jump sampling. These strategies may be used in a large variety of real-life inquiries. Looking to them generally researchers look for:

- Issues of sampling designs for vertices selection with particular properties,

- Unbiasedness of estimators of some graph property of interest.

Consider an undirected graph $G = (V, E)$. The vertex set is V and the edge set is E. In each vertex a function $f: V \rightarrow P$ is defined. For each problem is fixed the set P (set of the property values). The size of the graph is $M = |E|$ and the graph order is $N = |V|$. An edge between v and v' is denoted vv'.

An example is setting $P = \{days\ in\ hospitals,\ treatment\ received,\ sequel1, \ldots sequel\ k\}$. For a real problem, it is of interest the set of vertices $V^* = f^{-1}(P^*|P^* \subset P)$. P^* (target set) identifies a certain property described by values of interest. For example, P^* may be identifying the patients who had a particular treatment A. Hence $V^* = f^{-1}(P^*) = \{persons\ whose\ treatment\ was\ A\}$. In the sequel P* is assumed as independent of G. That is, the target set is drawn using a random sampling design from V and is fixed the cardinality $|V^*| = n^*$. In each concrete survey the expected search cost of finding a member of P^* depends only through n^*.

A naive approach to sampling epidemic graphs is the simple urn sampling procedure. It is very costly and inefficient but it is of use for evaluating the behaviour of other sampling designs. It works as follows:

- A sample of size n is selected from an urn
- n* <n are observed with certain property

Take the case for which the sampling allows the replacement of units. The number of draws $n^*(R)$ made before observing a first unit with the property is distributed according to a geometric distribution with probability of success. Probability $p=n^*/n$. Hence $E(n^*(R))= n/n^*$. In the no-replacement case the required random number of draws, $n^*(NR)$, has as expectation

$$E\left(n^*(NR)\right) = \frac{n+1}{n^*+1}.$$

The proof is derived in Stokes and Weber (2019).

The derived expected number of these two sampling design permits evaluating the behaviour using $\frac{E(n^*(NR))}{(n^*(R))} = \frac{(n^*+1)n}{n^*(n+1)} > 1$ unless $n^*=1$. Therefore, sampling without replacement is preferred because it improves the mean search time.

Consider some sampling methods for determining the behaviour of the dissemination of COVID-19, based in the fact that the population is described by a graph.

In epidemic studies a sample of vertices $s(V) = \{v_1, \dots, v_n\}, v_i \in V$ is selected using SRS. Consider a certain value of the property measured by $f(v_i)$.

In epidemic studies, using random sampling it is often unsuitable. Generally epidemic problems generate large size dynamic graphs. Therefore, it is needed to cope with the difficulties present when dealing with Big Data. This is due to the requirements on the order/size and/or and the evolution of the graph structure. So, there is a need to hold the graph in local memory. That is the reason for considering better alternatives to those using the graph structure. Commonly it is considered selecting a random sample of persons identifying them with vertices.

A goal of the sampling research is to obtain a representative subset of vertices. This is the usual motivation when identifying people with asymptomatic COVID-19. Sampled vertices are people. The sample may be obtained directly from health records, using a random street survey, etc. But in epidemics the target population is hidden and the researchers execute a sampling algorithm on a graph to identify the hidden population. For example, the selected individuals are tested for COVID-19 using a PSR test.

The objective in sampling graphs is to estimate its properties. Hence the researcher expects that the property is preserved on G_s. The calculated f (G s) is an estimator of f(G). If the need is only estimation, preservation of the properties on G_s are no so important. Knowing that the properties are biased, a correction is possible.

Though the motivation of property estimation and property preservation look different initially, they are closely related and can sometimes be transformed to each other. Consider the original graph G and sampled graph G_s.

When the cuts are preserved on G_s, weights of the cut may be evaluated and used as a naive and natural estimator for it in G. In the opposite case, cuts are not preserved on G_s – G_s should be transformed, so that the cuts are preserved. For example, if the interest is on the connections among people and the edges are sampled with fixed probability p, naturally the cuts of G_s have less edges than G. In up-weighting all the edges in G_s by a factor of p^{-1} the weighted cuts are preserved.

6.3. Remarks

- All property preservation results lead to good property estimators.
- Not all property estimation results preserve the property results.
- Directing property estimation straightly, without deriving previous property preservation result for mere estimation purpose, simplifies the inference.

7. Some sampling procedures

See Lee-Kim-Jeong (2006), Zhang-Patone (2017), Zhang -Ŏguz-Alper (2020), Leskovec- Faloutsos (2006) and Lewis (2011) for a discussion on sampling in networks. A discussion is given below on some common procedures for determining graph samples.

7.1. q- order Graphs

For organizing some ideas in the sequel, it is convenient to consider the order of the graphs.

7.1.1. First-order graph

First-order graph identifies the vertices without considering the edges. For example, a population $V = \{v_1, \dots, v_N\}$ is considered and a variable

$$q_i = \begin{cases} 1 \; if \; v_i \; holds \; the \; characteristic \; of \; interest \\ 0 \; otherwise \end{cases}$$

Note that the evaluation of a population U of people permits to determine the set.

$$V = \{v_i \in U | q_i = 1\}, |V\ | = \sum_{i \in U} q_i$$

taking

$$q_i = \begin{cases} 1\ if\ v_i\ is\ positive\ to\ a\ PSR\ test \\ 0\ otherwise \end{cases}$$

Therefore, the set of individuals with COVID-19 virus (a hidden population) is identified once a census of U is evaluated.

7.1.2. Second-order graphs

The set of the pair of nodes determines a second-order graph. Each pair is called a dyad. Take the set of dyads $V(2) \subset V$

$$V(2) = \{u, v | u, v \in V\}.$$

The dyads are of use for determining the size of G by selecting a sample form V(2) as well as for estimating the number of adjacent nodes and of mutual relationships. See a complete discussion on the inferences based on dyads sampling in Frank (1971, 1980).

Dyads are subgraphs of size 2 consisting of a pair of nodes and all the existing ties between them. Take a dyad determined by an unordered pair of nodes. The edge that exists between them may be characterised by $e_{ij} = (a_{ij}, a_{ji})$, the pair of binary variables. Then e_{ij} has four states in dependence of the values of the variables. In the sequel, it is assumed that the sampling design is Simple Random Sampling. There are three dyadic isomorphism states or classes:

Mutual relationships ($e_{ij} = (1,1)$)

$MR = \sum_{i,j} MR_{ij};\ MR_{ij} = \begin{cases} 1\ if\ e_{ij} = e_{ji}(1,1) \\ 0\ otherwise \end{cases}$ by

Take $P_{MR} = P(MR_{ij} = 1)$. An adequate estimator is

$$\widehat{MR} = \frac{N(N-1)}{n(n-1)} \sum_{i,j \in G_s} MR_{ij} = N(N-1)p_{MR};$$

its Mean Squared Error is

$$V(\widehat{MR}) = \frac{\left(N(N-1)\right)^2}{n(n-1)} \quad P_{MR}(1-P_{MR})$$

Asymmetric dyads ($e_{ij} = (1,0)\ or\ e_{ij} = (0,1)$)

$$AS = M_d - 2MR = \sum_{i,j} AS_{ij};\ AS_{ij} = \begin{cases} 1\ if\ e_{ij} = (1,0) or\ e_{ji}(0,1) \\ 0\ otherwise \end{cases};\ M_d$$
$$= \text{number of edges in the digraph;}$$

Denote $P_{AS} = P(As_{ij} = 1)$ an unbiased estimator is

$$\widehat{AS} = \frac{N(N-1)}{n(n-1)} \sum_{i,j \in G_s} AS_{ij} = N(N-1)p_{AS}$$

The Mean Squared Error is the variance

$$V(\widehat{AS}) = \frac{\left(N(N-1)\right)^2}{n(n-1)} P_{AS}(1-P_{AS})$$

Null dyad ($e_{ij} = (0,0)$)

$$ND = \frac{N(N-1)}{2} - MR - AS = \sum_{i,j} ND_{ij}, ND_{ij} = \begin{cases} 1\ if\ e_{ij} = e_{ji}(0,0) \\ 0\ otherwise \end{cases};\ N$$
$$= number\ of\ nodes\ in\ the\ digraph\ MR$$

Similarly denoting $P_{ND} = P(MN = 1)$

$$\widehat{ND} = \frac{N(N-1)}{n(n-1)} \sum_{i,j \in G_s} ND_{ij} = N(N-1)p_{ND}$$

and

$$V(\widehat{ND}) = \frac{\left(N(N-1)\right)^2}{n(n-1)} P_{ND}(1-P_{ND})$$

A sketch of the relationships in digraphs (directed graphs) is the following:

$v \leftarrow u\ (Asymmetric\ dyad)\ \ v \rightarrow u\ (Asymmetric\ dyad)$
$v \leftrightarrow u\ (Symmetric\ dyad)\ \ v \sim u\ (Null\ dyad)$

A key question in COVID-19 research is related with directed graph probability distributions. Fixing an adequate distribution allows testing hypotheses about properties of a directed graph. In epidemic studies it is needed to evaluate a hypothesis on the number of certain dyads, take for example $|\{(u,v)|u \rightarrow v; u\ infected\}|$.

7.2. Theoretical probability distributions

In practice simple distributions are usually assumed. Take $G_N(V)$ as the set of all possible labelled and irreflexive directed graphs, where $N = |V|$. Some of them are.

- Uniform distributions

$$P\big(I(e_{ij} = 1\big) = \frac{1}{2^{N(N-1)}}$$

- Bernoulli distribution

$$P\big(I(e_{ij} = 1\big) = \begin{cases} P_{ij}\ if\ i \neq j \\ 0\ otherwise \end{cases};\ P_{ij} \in [0,1]$$

When analysing *G,* the epidemiologist considers that a random phenomenon generates the edges. Hence even *G* may be considered as a random graph. Assuming independence

$$M_d = \sum_{ij} I(e_{ij});\ I(e_{ij}) = \begin{cases} 1\ if\ exists\ an\ edge\ between\ i\ and\ j \\ 0\ otherwise \end{cases}$$

It is distributed according to the binomial distribution $B(N(N-1), P)$. Hence

$$E(M_d) = N(N-1)P, V(M_d) = N(N-1)P(1-P)$$

In different studies, a sample of dyads is selected by the epidemiologists for measuring the degree of reciprocity in a network. That issue is of interest

when analysing the contacts among people and the possible exchange of fluids, for example.

Analysing *G*, a dyad census may be developed and the different types of dyads that can occur are identified. Then, the expected numbers of the different types of dyads under some specific distribution may be computed. A sample graph is generated for evaluating the effect of their expectations using appropriate statistical method, having *G* simulations may be performed for evaluating the effect of policies in the dynamics of COVID-19 pandemic. Similarly, tests of hypotheses on the number of choices and of mutual choices for a specific relation may be performed. In COVID-19 studies, the epidemiologists are interested in estimating the spread of the virus, estimating the number of mutual contacts of contaminated people (specific relation) etc. Also, to establish if the expectation has a certain relationship with a pre-specified value is among the problems to be answered using a sample graph. With the sample information obtained of n(n-1) dyads, the parameters of interest may be estimated using the above-presented estimators. Statistical inferences may be performed using the properties of statistical estimation theory. Generally, it is assumed that the sample size is large enough for using the normal approximation of the Binomial distribution.

7.3. Third-order graph

Third-order graphs are determined by three nodes (triad). Denote the corresponding set $V(3) \subset V$ by

$$V(3) = \{u, v, w | u, v, w \in V\}.$$

Triads are subgraphs of size 3 formed by 3 nodes, $i \in V, j \in V, k \in V$ and the edges among them. They are of use for determining the size of G by selecting a sample from *V(3)*. In particular they are of interest in some applications. Take for example the interest triangles. They are used for considering equivalence relations. The number of three nodes forming a triangle, in an undirected graph, determines issues on the transitiveness of the relationship. Frank (1981) established the existence of an explicit relationship between the mean and variance of the degree distribution and the triads of a graph.

Once the different kinds of triads are counted, some network`s properties of G may be studied. Examples: the balancing, clustering issues, transitivity, microstructures exhibited in patterns etc.

Denote a triad by T_{ijk}, $i < j < k$. There are six possible ordered triples for each triad and there are sixteen triad types ($1 \leq u \leq 16$). In G there are $N_T = \frac{N(N-1)(N-2)}{6}$ possible triads. The triads are not independent. Take

$$l_u = \textit{the coefficients of the linear combination } u$$

$$\binom{N}{3} = \sum_u T_u = \textit{The counts in the triad census}$$

$M_d = \frac{1}{N-2}\sum_u c_u T_u = \textit{The counts of edges in a census},$

c_u = the number of edges in the u-*th* triad type

Performing a triad census in G is possible computing the linear combinations

$$\tau = \sum_u l_u T_u$$

Under the assumed model, the digraph is random. The probabilities involved are given below:

- The probability that the θ-subgraph K* belongs to the isomorphism class C(u) $= P(K^* \in C(u))$.
- The probability that a pair of the k-subgraphs $K^* \in C(u), U^* \in C(v)$ belongs to given classes $= P(K^* \in C(u), U^* \in C(v))$.
- The average of the probabilities $P\left(K^* \in C(u)\right) = \bar{P}(u) = \frac{1}{\binom{N}{\theta}}\sum_{K^*} P\left(K^* \in C(u)\right)$.

- The joint probability of observing two classes= $\bar{P}_j(u,v) = \frac{1}{\binom{N}{\theta}\binom{N-\theta}{\theta-j}\binom{\theta}{j}} \sum_{K^*,U^*} \quad P(K^* \in C(u), U^* \in C(v))$.

Denoting N(u) as the number of θ-subgraphs in C(u) its expectation is:

$$E\big(N(u)\big) = \binom{N}{\theta}\bar{P}(u)$$

The variance and covariances are:

$$V\big(N(u)\big) = \binom{N}{\theta}\Bigg\{\bar{P}(u)\big(1-\bar{P}(u)\big) + \sum_{j=0}^{\theta-1}\binom{N-\theta}{\theta-j}\binom{\theta}{j}\Big[\bar{P}_j(u,u) - \big(\bar{P}\ \ (u)\big)^2\Big]\Bigg\}$$

$$COV(N(u),N(v)) = \binom{N}{\theta}\Bigg\{-\bar{P}(u)\big(\bar{P}(v)\big) + \sum_{j=0}^{\theta-1}\binom{N-\theta}{\theta-j}\binom{\theta}{j}\Big[\bar{P}_j(u,u) - \big(\bar{P}\ \ (u)\big)^2\Big]\Bigg\}$$

From these formulae the parameters of T_u are derived as,

$$E(T_u) = \binom{N}{3}\bar{P}(u)$$

$$V(T_u) = \binom{N}{3}\Bigg\{\bar{P}(u)\big(1-\bar{P}(u)\big) + \sum_{j=0}^{2}\binom{N-3}{3-j}\binom{3}{j}\Big[\bar{P}_j(u,u) - \big(\bar{P}\ \ (u)\big)^2\Big]\Bigg\}$$

$$COV(T_u, T_v)) = \binom{N}{3}\Bigg\{-\bar{P}(u)\big(\bar{P}(v)\big) + \sum_{j=0}^{2}\binom{N-3}{3-j}\binom{3}{j}\Big[\bar{P}_j(u,u) - \big(\bar{P}\ \ (u)\big)^2\Big]\Bigg\}$$

Note that τ is a linear combination, then linear estimation results are usable and

$$E(\tau)=\binom{N}{3}\ \sum_u\ \ l_u\bar{P}(u)$$

$$V(\tau)=\sum_u\ \ l_u^2\left\{\bar{P}(u)\big(1-\bar{P}(u)\big)+\sum_{j=0}^{2}\binom{N-3}{3-j}\binom{3}{j}\Big[\bar{P}_j(u,u)-\big(\bar{P}\ \ (u)\big)^2\Big]\right\}-\binom{N}{3}\sum_{u\neq v}l_u l_v\bar{P}(u)\big(\bar{P}(v)\big)+\binom{N}{3}\sum_{u\neq v}l_u l_v\sum_{j=0}^{2}\binom{N-3}{3-j}\binom{3}{j}\Big[\bar{P}_j(u,u)-\big(\bar{P}\ \ (u)\big)^2\Big]$$

The distributional properties of linear estimators allow accepting, in certain cases, that multivariate statistical inference may be performed under the validity of approximate multivariate normality. Hence, the involved probabilities are estimated using the observed frequencies and the "plug-in-rule". The consistency of the corresponding linear estimators may support normal inferences validity.

Structural hypotheses may be tested using triads. Ideas on the configurations of a subset of nodes and arcs may be translated into mathematical statements about triads. The epidemiologist has ideas on the effect of configurations occurrence.

Zhang-Pattone (2017) established analogies that unified the treatment of sampling of dyads and triads as well as the subpopulation graph sample.

7.4. Sampling vertices

Sampling dyads or triads directly represents insurmountable logistic costs to epidemiologic research. It seems better to select some vertices and construct a graph using them as seeds. See detailed discussions in Ahmed et al. (2011). The basics on sampling vertices are discussed below.

7.4.1. Vertex Sampling

- **Vertex Sampling (VS)** is a class of sampling procedures where a sample s is selected from V and $V_s \subseteq V$ is obtained uniformly or according a specific distribution. Usually, the epidemiologist uses only information associated with the vertices, as the interest is interviewing people and determining relations and other issues.

- **Vertex Sampling with Neighbourhood (VSN)** is a variant VS. A sample of vertices $V_s^* \subseteq V$ is selected directly without information on the topology. Now, $E_s = \left\{\bigcup_{v \in V_s^*} \delta(v)\right\}$ and $V_s = \{v', v | (v', v) \in E_s\}$. The sampled graph is $G_s = (E_s, V_s)$. When interested in the crawling this procedure is more realistic than VS.

VS and VNS are simple and the epidemiologists may use the sampled graphs for developing theoretical analysis on the behaviour of the pandemic. The selected samples using these methods are uncorrelated.

- **Random Star Sampling (RSS).**

Take the definition of a star sample of a graph:

Definition. $N_{Ge}(v)$ is a star sample with star center v and star points $N_G(v)$.

Each v_i determines the centre of a star (CS). A direct neighbourhood of a vertex v is determined by identifying the vertices directly connected with it: $N_G(v) = \{v' \in V | (v', v) \in E\}$. An extended neighbourhood is $N_G(v) = \{v' \in V | (v', v) \in E\} \cup \{v\}$. The corresponding edge neighbourhoods are $\Gamma_G(\mathrm{v}) = \{v'v \in E\}$ and $\Gamma_{Ge}(\mathrm{v}) = \bigcup_{v' \in N_G(v)} \Gamma_G(v')$. The variables of interest are evaluated in the neighbourhood.

Note that $(v, v) \in E$ and the property $f(G_s)$ is evaluated. Take V*⊆ V and denote:

- $N_G(V^*) = \bigcup_{v \in V^*} N_G(v) \setminus V^*$, neighbours of V* not including V* .
- $N_{Ge}(V^*) = \bigcup_{v \in V^*} N_{Ge}(v)$, neighbours of V* including V*
- $d_G(v) \equiv |N_G(v)|$ the degree of v
- $de_G(v) \equiv |N_{Ge}(v)| = d_G(v) + 1$, the extended degree of v

- The set of vertices with degree k as $V_G(k) = \{v \in V | d_G(v) = k\}$
- A by-degree partition of V into subsets $V_G(k)$, $k \in D$.
- $w_G(k) = \frac{|V_G(k)|}{n}$ as the fraction of vertices with degree k

The degree distribution of G is $\{w_G(k)|k \in D\}, \sum_{k \in D} w_G(k) = 1$. Therefore the expected degree of v is:

$$E(d_G(v)) = \frac{1}{|V|}\sum_{v \in V} d_G(v) = d_G, v \sim (V)$$

This is a mean and an appropriate estimator is the sample mean

$$\hat{d}_G = \frac{1}{|V_s|}\sum_{v \in V_s} d_{G_s}(v)$$

It is normally distributed for a sufficiently large sample size. Hence $\hat{d}_G \sim N\left(d_G, \frac{\sigma_G^2}{|V_s|}\right)$ and normally based inferences may be performed. Note that then is accepted:

$$\hat{\sigma}_G^2 = \frac{1}{|V_s| - 1}\sum_{v \in V_s}\left(d_{G_s}(v) - \hat{d}_G\right)^2$$

It is an unbiased estimator of the variance.

RSS provides a framework for identifying possible infected people. Once an infected person is detected in the sample in pandemic studies or for clustering similar recovered people in a community, the neighbourhood may be easily obtained by interviewing the patients. Sampling should consider in the designing stage some cost model. The models rely on the number of star samples, considering the cost per unit, or on the number of vertices in relation to the values of the property. The first model is used when queries' cost incurred is computed from the star and the property values of the vertices in its neighbourhood. The second one identifies costs incurred by the computation of the property values.

The selection of the nodes may be developed using one of the sampling designs available in the literature. Some popular sampling designs used in applications are described below.

- RSS with replacement (SRSWR): The same probability is assigned to each vertex and $s(V) = \{v_1, \dots, v_n\}$ is determined using SRS with replacement. The n selected vertices are the star centers.

This procedure generates an iid sequence of random SC's (v_t , $t \in N$), with $v_t \sim uni(V_0)$. The procedure allows replacement, hence $G_t = G_0$ for any t.

- RSS without replacement (RSS-WSR): The same probability is assigned to each vertex and $s(V) = \{v_1, \dots, v_n\}$ is determined using SRS without replacement of SC's. In each stage, the drawn SC are removed from the graph in the next selection.

This sampling determines a sequence of SC's (v_t , t [n]), with $v_t \sim \text{uni}(V_{t-1})$. The graph is updated and $V_t = V_{t-1} \setminus \{v_t\}$, and $E_t = E_{t-1} \setminus \Gamma_{G_{t-1}}(v_t)$.

- RSS without star replacement (RSS-WOSR): A first SC is selected from V. The centre and all the adjacent edges are removed from the graph after the query. The procedure is applied n times.

The random sequence of star centres generates (v_t , t {1,...,T}), with $v_t \sim uni(V_{t-1})$, and the graph is updated by removing the star and the edges in its extended edge neighbourhood. Therefore

$$V_t = V_{t-1} \setminus N_{Ge_{t-1}}(v_t) \text{ and } E_t = E_{t-1} \setminus \Gamma_{Ge_{t-1}}(v_t).$$

See that using RSS for sampling a graph, $G = (V, E)$ generates a sequence of graphs, $G_t = (V_t, E_t), t \in$ and of a vertex sets $V_t, t \in$. Both the vertex and edge sets are random variables. Take $G_0 = (V_0, E_0)$ as the initial random graph and random target set V_0^*, corresponding to t = 0. Note that $if\ V_t^* = V_t \cap V_0$ holds the units in the initial target set are included into the t samples. Note that the procedure described ensures that

$$V_t \subseteq V_{t-1}, E_t \subseteq E_{t-1}, V_t^* \subseteq V_{t-1}^*.$$

Stokes and Weber (2019) developed a study of the costs of these selection procedures. Considering the cost, the accuracy approximation to them is very good for SSRWOSR. The probability of selecting a certain star at

selection *t*, conditioned to not being selected previously, may be approximated. The common approach consists in evaluating in the previous *t-1* samples the amount of the expected reduction in the overall order of the graph and in the number of neighbours of the target set. This yields an approximate unconditional probability of first hitting the target set at *t*, which in turn yields an approximate expected cost.

The fact is that in epidemiologic studies, large graphs appear, which are encoded in some ways. The structures of the data attached to each vertex indicate the profile of a particular person (patient), and allow identifying as in the neighbourhood of other people connected in a certain network. In COVID-19 pandemic studies, connections of infected people with others are of interest for identifying the spread of the virus. Post-pandemic, selecting recovered patients and identifying others with similar conditions (treatments, hospital, etc.) allows the development of studies based on sequels after recovering.

7.4.2. Traversal-Based Sampling

From the limitations present in node-based and edge-based sampling methods, a third category of sampling models was developed. It is called traversal-based sampling (TBS). This class of sampling methods is very old, but only recently graph research has focused on TBS methods – see Leskovec and Faloutsos (2006). They are identified also as topology-based sampling or sampling by exploration methods. These methods maintain the connectivity of the graphs after sampling. This class of sampling methods are based on the selection of a set of initial vertices. The sample is expanding based on these observations. Note that this is the underlying idea on many epidemiological studies on the spread of the virus. That is the goal of pandemic research, as the main interest is detecting hidden infected people. In most real applications, one cannot perform VS and ES directly due to different constraints. Among them is that to enumerate it is too costly. In such cases, TBS becomes more practical, as it only relies on a small sized starting topology (a few seed nodes). From the initial seeds, the sample is expanded during the exploration.

Note that VS, ES and TBS share ideas and algorithms. When necessary, certain TBS techniques may be used as generators in VS or ES. In addition, if TBS is used for mimicking VS or ES and the number of probes is increased substantially, samples generated using TBS have correlations. This fact may be utilized for improving the efficiency of estimators, as some

clustering coefficient. Real life experiences recommend using TBS directly as it is less costly than mimicking VS and ES.

Some particular methods are described below.

- First (BF), Depth First (DF) and Random First (RF) sampling are particular algorithms of TBS sampling. A randomly chosen node is a seed and a set of connected elements is selected.

The method called Breadth-First Sampling (BFS) is a node-sampling algorithm. It has been widely studied – see Lee et al. (2006), Kurant, et al. (2010, 2011) and Wilson (2009). Once a seed v is selected, BFS finds the nodes that are close to it and determines distance for graph analysis. It works as follows: the first node v in the queue "Processed" is sent to queue "Sampled". The neighbours of node v are inserted into the queue "Processed", unless it has been previously processed. Another randomly selected node is stored in "Processed". The process continues until the fixed goal (budget) is obtained. BFS is biased towards nodes with high degree. That is, BFS, nodes with a higher degree, are visited more frequently determining high local clustering coefficients due to its bias. Algorithmically it is described as follows:

Fix B (Budget) , $Q \leftarrow \{v_0\}$, $V_s \leftarrow \{v_0\}$, $L = \{\}$.
Dequeue $v = Q.dequeue()$
$B \leftarrow B - b$, $L \leftarrow L + \{v\}$
For $u \in N(v) \cap u \notin L \cap u \notin Q$ invoke $Q.enqueue(u)$

Then V_s =L is the sample of vertices; the edges are determined identifying the connexions of the nodes in L.

DFS and RFS may be described similarly. The difference among them is in the implementation of the function "dequeue ()". In BFS, the first node is selected; in DFS, the last node is selected; in RFS, a random node is selected.

- Snow-Ball sampling (SBS) is similar to BFS. The seminal paper is due to Goodman (1961). Some foundational ideas were given by Birnbaum-Sirken (1965). More details may be obtained in Frank-Snijders (1994). BFS includes all the neighbours of a selected vertex, while SBS selects only a fixed fraction. The method exhibits a boundary bias that peripheral node miss a number of its neighbours.

It is popularly used in investigations on hidden population. As quoted, epidemic researches deal with such populations. Note that epidemiologists need to detect and study asymptomatic infected people. This method is also called "network sampling" or "chain referral sampling" – see Sirken (2005) and Blagus et al. (2015). It works as follows:

Select an initial sample of seeds $s_{1,t}$

Initialize T, t=0,

For t<T

$$Successors(s_{1,t}), t = t + 1$$

$$s_{1,t} = Successors(s_{1,t-1}) \setminus \bigcup_{h=0}^{t-1} s_{1,h}$$

If $s_{1,t} = \phi \; then, for \; h > t \; , s_{1,h} = \phi$

Else t=t+1

Each loop determines a wave.

- Forest Fire Sampling (FFS) is a probabilistic extension of SBS. The method starts with the selection of a seed node. Its incident edges and adjacent nodes are eliminated (burned) recursively. After burning the edges, the endpoints are collected and the process is repeated until sufficient nodes have been visited. The process relies on the generation of k independent Geometric variables with parameter *p=1/k*. FFS model captures some important issues when dealing with social networks. Therefore, in post pandemic studies, it is efficient as an infected person drives to a set of reassigned contacts. The use of FFS for estimation behaves as SBS. FFS is not seriously biased to high-degree nodes, as it avoids the selection of previously traversed nodes.

A number of studies have evaluated the performance of these algorithms by measuring the properties of sampled graphs – see for example Ahmed et al. (2011, 2012).

8. Motivations for using sampled graphs in COVID-19 studies

Graph sampling is a set of statistical models for studying real life graphs. Graphs are commonly used to represent the structure of many phenomena

from engineering, social, medical or biological phenomena. The variation over all the possible subsets of nodes and edges of a population graph is of interest. Sample graphs are taken from a given population graph, using some sampling method. Zhang and Patone (2017) discussed at large on graph sampling theory. Initial seminars, papers and ideas are due to Frank (1971, 1980 a, b). See also Frank (2011), Bouza – Allende (2002) and Bouza – Allende-Negreiros (2015). The last decades have witnessed an explosive growth in social networks. Their size is huge and researchers cope with the need of better understanding these graphs. Acquiring the complete view of a graph representing a network is rather impossible, and surely too costly. Then, the theme of sampling graphs is attracting the attention of investigators of different areas. Epidemic problems may be studied using the graph generated by patients, contacts, hospitals etc. When coping with huge size network graphs, researchers tend to use, for a better understanding, a sample of these graphs. Several sampling methods and algorithms have been proposed in the literature – see Frank (2011).

A network is a valued graph. The structure of a network is determined by a collection of nodes and edges joining the nodes. Measures may be attached to the nodes or to the edges (or both) and a valued graph is determined. The research may consider modelling the population network as a random realisation, or to base the modelling in the possible samples of a given fixed population network. Graph sampling is concerned with the structure of a network and its perspective – see discussions in Goldenberg et al. (2010).

Graph sampling is a particular problem within the theory of probability sampling. A general definition is given for probability sample graphs, in a way that it is similar to general probability samples from a finite population. In practice, it is hard to acquire the complete graph in many real-life networks. The researcher commonly looks on a given graph for characteristics, summarized by a parameter $\eta(G)$, as the order, size, degree, clustering coefficient etc. In real life, it is impossible or costly to calculate $\eta(G)$ from the population graph G. The sample graph G_s permits the estimation of it by evaluating $\hat{\eta}(G_s)$, a plug-in estimator. It often provides a poor representation of the parameter.

Example 1.

Let $G(V,E)$ be a network of people recovered from COVID-19. The epidemiologist is interested in the average degree

$$\eta(G) = \frac{1}{|V|}\sum_{v \in V} d(v), d(v) = degree\ of\ v$$

Selecting a sample of vertices $V_{s,}$ a naïve estimator is the sample mean.

$$\hat{\eta}(G) = \frac{1}{|V_s|}\sum_{v \in V_s} d(v), d(v) = degree\ of\ v$$

In the inference process it is needed to incorporate the effects of random sampling; and/or of measurement errors. The sampling design, the topology of *G* and the nature of *η(G)* play a key role in the generalizations to be made from G_s to *G*. The researcher may decide to use Model-based inference, and use Likelihood-based or Bayesian paradigms, or a Design-based method, and use Statistical Sampling Theory tools, assuming that observations are made without measurement error.

9. Numerical experiments

A census was performed for determining the relationships among a population. The experiment was performed in a medium-size village (5,541 inhabitants). Social workers censused and determined the contacts among villagers. Connections with outsiders were very reduced, and hence not included. Students of primary and secondary schools received classes in installations of the village. The villagers were mainly farmers and artisans. They were not moving to next cities and towns with frequency. Then the movements to other places were scarce. The data collected has allowed determining a network *G*. Infected people in the community were considered as the source of the spread of the virus.

9.1. Analysis based on dyads

The number of pairs of nodes in *G* was *N(N-1)=30 697 140*. An experiment was performed and *10,000* sample graphs of corresponding sizes B=*30 000, 3 000* and *300* were generated. In each, G_s were estimated MR, AS, ND and M_d. The variables $\widehat{MR}, \widehat{AS}, \widehat{ND}\ and\ \widehat{M}_d$ were distributed Binomial with probabilities of success $P_Q, Q = MR, AS, ND, M_d$. The estimation of them was indexed by the graph corresponding sampled graph as $p_Q(G_s), Q = MR, AS, ND, M_d$

Due to the knowledge of *G,* these parameters are known, and it is possible to evaluate the behaviour of the sample graphs computing for each sample size the Relative Mean Absolute Error (MAE)

$$\varepsilon\left(\hat{Q}\ \ (B)\right) = \frac{1}{10\,000}\sum_{s=1}^{10\,000}\frac{\left|\hat{Q}\ \ (G_s) - Q\ \ \right|}{Q},$$

$$Q\ \ = MR, AS, ND, M_d, B = 30\,000, 3\,000, 300$$

Note that a CLT may be valid for sustaining that

$$T(G_s) = \frac{\hat{Q}\ \ (G_s) - Q}{\sqrt{\frac{N(N-1)p_Q(G_s)(1-p_Q(G_s))}{|V_s|(|V_s|-1)}}}$$

is distributed approximately N*(0,1)* for a sufficiently large value of $|V_s|(|V_s|-1)$. Then the test of hypothesis

$$H\text{: } E\left(\hat{Q}\ \ (G_s)\right) = Q\ \ vs\ K\text{: } \left(\hat{Q}\ \ (G_s)\right) \neq Q$$

was made for each G_s and was computed

$$H(G_s) = \begin{cases} 1\ if\ H\ is\ accepted \\ 0\ otherwise \end{cases}$$

The procedure generates acceptable inferences when

$$\hat{\gamma} = \frac{1}{B}\sum_{s=1}^{B} H(G_s)$$

is close to $1-\alpha$, it was fixed in 0,95.

The experimental results in the simulated sampled graphs are given in the next table.

B	$\varepsilon(\widehat{MR})$	$\hat{\gamma}$	$\varepsilon(\widehat{AS})$	$\hat{\gamma}$	$\varepsilon(\widehat{ND})$	$\hat{\gamma}$	$\varepsilon(\widehat{M}_d)$	$\hat{\gamma}$
30 000	0,259	0,942	0,211	0,938	0,210	0,937	0,207	0,956
3 000	0,309	0,917	0,217	0,921	0,252	0,945	0,224	0,916
300	0,401	0,899	0,326	0,894	0,323	0,896	0,398	0,906

The sample sizes are relatively small compared with the number of possible pairs of villagers. The results suggest that the procedures are quite accurate even for the samples of 300 dyads. The tests based on 300 dyads are not so acceptable. In any case, the sampled graphs provide information for evaluating the dynamics of the epidemic graph at a low cost. In real life the parameters are not known, and the epidemiologist fixes a hypothetical value of Q considering a certain scenario.

10. Analysis based on triads

The number of triplets of nodes in G was N*(N-1)(N-2)=28 389 738 310*. Note that even in such a small population, analysing all the triads is unpractical. The experiment was performed and sample graphs of size 30 000, 3 000 and 300 were selected and the triads determined in each G_s.

$$\binom{n}{3} = \sum_u T_u(G_s) = The\ counts\ \ of\ the\ triads\ in\ G_s$$

Then

$$\widehat{M}_d = \frac{1}{n-2}\sum_u c_u T_u(G_s) \quad = The\ counts\ of\ edges\ in\ G_s,$$

c_u = the number of edges in the u-th triad type

and

$$\hat{\tau} = \sum_u l_u T_u(G_s)$$

Using the counts, are estimated

$$P(K^* \in C(u)), P(K^* \in C(u), U^* \in C(v)), \bar{P}(u)$$

By

$$p\big(C(u)\big) = \hat{P}(K^* \in C(u)) = \frac{number\ of\ \text{subgraphs classified in C(u)}}{number\ of\ \text{subgraphs}}$$

$$p(C(u)\&C(v)) = \hat{P}(K^* \in C(u), U^* \in C(v)) = \frac{number\ of\ \text{subgraphs classified in C(u)}\ and\ C(v)}{number\ of\ \text{subgraphs}}$$

$$\hat{\hat{P}}(u) = \frac{1}{\binom{n}{3}} \sum_{K^*} \quad p\big(C(u)\big)$$

$$\hat{\hat{P}}_j(u,v) = \frac{1}{\binom{n}{3}\binom{n-3}{3-j}\binom{3}{j}} \sum_{K^* \cdot U^*} \quad (C(u)\&C(v))$$

Hence

$$E\big(\widehat{T(u)}\big) = \binom{N}{3}\hat{\hat{P}}(u)$$

$$\hat{V}\left(\big(\widehat{T(u)}\big)\right) = \frac{\binom{N}{3}}{\binom{n}{3}}\left\{\hat{\hat{P}}(u)\left(1-\hat{\hat{P}}(u)\right) + \sum_{j=0}^{2}\binom{n-3}{3-j}\binom{3}{j}\left[\hat{\hat{P}}_j(u,v) - \left(\hat{\hat{P}}(u)\right)^2\right]\right\}$$

$$\widehat{COV}\big(\widehat{T(u)}, \widehat{T(v)}\big) = \frac{\binom{N}{3}}{\binom{n}{3}}\left\{-\hat{\hat{P}}(u)\hat{\hat{P}}(v) + \sum_{j=0}^{2}\binom{n-3}{3-j}\binom{3}{j}\left[\hat{\hat{P}}_j(u,v) - \left(\hat{\hat{P}}(u)\right)^2\right]\right\}$$

As a result

$\hat{\tau}=\binom{N}{3}\sum_u \quad l_u\hat{\bar{P}}(u)$

$V(\hat{\tau} \quad)= \quad \sum_u \quad l_u^2\left\{\bar{P}(u)\left(1-\bar{P}(u)\right)+\sum_{j=0}^{2}\binom{N-3}{3-j}\binom{3}{j}\left[\bar{P}_j(u,u)-\left(\bar{P} \quad (u)\right)^2\right]\right\}-\binom{N}{3}\sum_{u\neq v} l_u l_v \bar{P}(u)\left(\bar{P}(v)\right)+$
$\binom{N}{3}\sum_{u\neq v} l_u l_v \sum_{j=0}^{2}\binom{N-3}{3-j}\binom{3}{j}\left[\bar{P}_j(u,u)-\left(\bar{P} \quad (u)\right)^2\right]$

As quoted previously, multivariate statistical inferences may be performed under the assumption the validity of the normality. Then tests may be performed, using a constant value for l_u, by

$$T(\hat{\tau})=\frac{\hat{\tau}-\tau}{\sqrt{\hat{V}(\hat{\tau})}}$$

Computing it for each sampled graph, the evaluation of triads behaviour was done:

$$CV(G_s)=\frac{\sqrt{\hat{V}\left((\hat{\tau}|G_s)\right)}}{\tau}$$

The overall performance of the procedure was evaluated computing:

$$\varepsilon(\hat{\tau})=\frac{1}{B}\sum_{s=1}^{B} \quad \frac{\sqrt{\hat{V}\left((\hat{\tau}|G_s)\right)}}{\tau}$$

$$I[T(\hat{\tau}|G_s)]=\begin{cases}1 \; if \; is \; accepted \; E(\hat{\tau}|G_s)=\tau \\ 0 \; otherwise\end{cases}$$

was used, once the test of hypothesis was performed for each G_s for estimating the confidence coefficient $\gamma = 0{,}95$ by means of

$$\hat{\gamma}=\frac{1}{B}\sum_{s=1}^{B} \quad I[T(\hat{\tau}|G_s)]$$

The results are given in the table below:

B	$\varepsilon(\hat{\tau})$	$\hat{\gamma}$
30 000	1,006	0,871
3 000	1,764	0,855
300	3,241	0,704

For triads, the results are not so accurate. The cause may be due to the fact that the asymptotic normality may not correct even when B=30 000 triads. It seems that it is not large enough for supporting the validity of the convergence linear estimator.

11. The behaviour of cut set measures

As discussed previously, the cut of a set S, $Cut(S) = |\delta(S)| = |\{(u,v) \in E | u \in S, v \notin S\}|$, is the number of edges crossing S. Assuming that S is generated by a SRS mechanism, the epidemiologist is able to evaluate the number of people contaminated in S contacting the remaining population, within a statistical framework. The impact in infected people, possibly due to transmission from S, in the remaining uncontacted people, may be evaluated. Once a group of people S is tested for COVID-19, it is important to evaluate the impact of the spread of the virus in the remaining population. That is important in designing pandemic policies and evaluating the efficiency of management protocols.

Note that

$$RCut(S) = \frac{|\delta(S)|}{|S|}$$

is a proportion in the sample,

$$Ncut(S) = \frac{|\delta(S)|}{|vol(S)|}$$

And

$$Cut(S) = \frac{|\delta(S)|}{\sum_{v \in S} w(v)}$$

are ratios. Denote them by

$$p(RC), RC = \frac{|\delta(S)|}{|S|}, \frac{|\delta(S)|}{|vol(S)|}, \frac{|\delta(S)|}{\sum_{v\in S} w(v)}.$$

These measures are computed in the sample of vertices. $|\delta(S)|$ is a binomial random variable with parameters $|S|$ and $P_\delta = Prob((u,v) \in E)$. Their probabilistic behaviour is described by the distribution $Bin(n^*, P_\delta)$, $n^* = |S|, |vol(S)|, \sum_{v\in S} w(v)$. The CLT supports the normal approximation of

$$T(RC) = \frac{p(RC) - P_{\delta 0}}{\sqrt{p(RC)(1 - p(RC)/n^*}} \sim N(0,1)$$

The epidemiologist may test if the hypothetical value of P_δ , fixed as $P_{\delta 0}$, is valid. That is accepting $H: P_\delta = P_{\delta 0} or\ K: P_\delta \neq P_{\delta 0}$. In the experiment we know $\delta_0 = \delta$.

The accuracy of the measures is evaluated by computing the Relative Mean Absolute Error (MAE)

$$\varepsilon\big(T(RC)\big) = \frac{1}{10\ 000} \sum_{s=1}^{10\ 000} \left| \frac{p(RC) - P_{\delta 0}}{P_{\delta 0}} \right|_{G_s}$$

The results of the simulation are presented in the following Table:

	$RCut(S)$	$Ncut(S)$	$Cut(S)$	$RCut(S)$	$Ncut(S)$	$Cut(S)$	$RCut(S)$	$Ncut(S)$	$Cut(S)$
		3000			300			50	
VS	1,51	1,61	1,40	1,61	1,50	2,10	1,66	1,61	1,57
VSN	1,50	1,91	1,51	1,81	1,45	1,09	1,99	1,40	1,78
RSS-WR	1,11	1,21	1,10	1,31	1,11	1,09	1,31	1,11	1,17
RSS-WOR	0,91	1,01	0,97	1,11	1,01	0,93	1,11	1,09	1,21
BFS	0,91	0,95	0,94	0,94	1,09	0,92	1,04	0,91	0,91
RFS	0,88	0,98	0,91	0,94	1,01	0,91	0,96	0,99	1,08
SBS	1,18	1,51	0,97	0,98	1,33	0,93	1,36	1,39	1,51
FFS	0,98	1,61	0,97	1,72	1,28	1,07	1,18	1,24	1,49

Table: Mean Absolute Error of cut measures

The MSE's are stable for the changes in B. The more accurate sampling methods were BFS and RFS. VS and VSN exhibited large values of the MSE in all the cases.

The test was developed for each sampled graph and was computed

$$\hat{\gamma}_{RC} = \frac{1}{B}\sum_{G_S} I[T(RC)]_{G_S};\ I[T(RC)]_{G_S} = \begin{cases} 1\ if\ H\ is\ accepted \\ 0\ otherwise \end{cases}$$

	$RCut(S)$	$Ncut(S)$	$Cut(S)$	$RCut(S)$	$Ncut(S)$	$Cut(S)$	$RCut(S)$	$Ncut(S)$	$Cut(S)$
		3 000			300			50	
VS	0,94	0,93	0,91	0,93	0,90	0,80	0,86	0,82	0,75
VSN	0,77	0,75	0,72	0,75	0,80	0,77	0,70	0,71	0,70
RSS-WR	0,93	0,95	0,87	0,93	0,85	0,88	0,89	0,82	0,74
RSS-WOR	0,84	0,79	0,72	0,83	0,83	0,84	0,74	0,77	0,70
BFS	0,75	0,76	0,69	0,72	0,70	0,68	0,58	0,66	0,64
RFS	0,68	0,68	0,61	0,69	0,71	0,59	0,63	0,60	0,53
SBS	0,68	0,64	0,57	0,68	0,67	0,63	0,62	0,55	0,51
FFS	0,78	0,62	0,67	0,64	0,65	0,67	0,70	0,54	0,52

Table: Estimated $\gamma\ for\ 1-\alpha = 0{,}95$ in tests on cut measures

The results of Vertex sampling, described in section, suggests that VS and RSS-WR have a good behaviour, as the test's level is close to 0,95. RSS-WOR seems to be acceptable for n*>300. VSN has the worst behaviour. Note that for *n*=50,* the tests have considerably smaller values of $\hat{\gamma}$. The results for Traversal sampling were worse than those obtained for vertex sampling. In all the cases, the estimated $\hat{\gamma}$ is considerably far from 0,95.

12. Path measures

$\widehat{GD} = N\sum_k kp_{Deg}(k)$ is the estimator of the Graph Density and testing hypothesis on the graph density described by

$$H: GD = N\sum_k kP_{0Deg}(k)\ vs\ K: GD \neq N\sum_k kP_{0Deg}(k)$$

It may be developed using the fact that the distribution of

$$T_{GD} = \frac{\widehat{GD} - GD}{\sqrt{N^2 \sum_k k^2 \frac{p_{Deg}(k)(1 - p_{Deg}(k))}{|V_s|}}}$$

is approximated by a N(0,1).

For the Average Path Length, the proposed estimator and the estimated error are:

$$\widehat{APL} = \frac{1}{\binom{|V_s|}{2}} \sum_{1 \le i < j \le |V_s|} \quad arg\ min_t [P_t]_{i,j}; S(\widehat{APL}) = \frac{1}{\binom{|V_s|}{2}} \sum_{1 \le i < j \le |V_s|} \quad (arg\ min_t [P_t]_{i,j} - \widehat{APL})^2$$

It seems that may be valid accepting the normality of $T_{APL} = \sqrt{\frac{\widehat{APL} - APL}{S(\widehat{APL})}}$;

For ratios and diameters, the estimation considered used

$\hat{\bar{R}} = \frac{1}{|V_s|} \sum_{v \in V_s} \quad R(v)$ and $\widehat{\overline{DG}} = \frac{1}{|V_s|} \sum_{v \in V_s} \quad DG(v)$.

It means, both estimators may be considered as convergent to a normal distribution under large value of $|V_s|$ and the infrences will be based on the T-Student statistic.

$$T_{PM} = \frac{\widehat{PM} - \overline{PM}}{S(\widehat{PM})}; PM = R, DG, s\left(\widehat{PM}\right) = \sqrt{\frac{1}{|V_s|(|V_s| - 1)} \sum_{v \in V_s} \quad \left(R(v) - \hat{\bar{R}}\right)^2 \bar{\varepsilon}(Y)} = \frac{1}{|V_s|} \sum_{v \in V_s} \quad \varepsilon(Y(v)), \varepsilon(Y(v)) = \varepsilon\left(\widehat{DG}(v)\right), \varepsilon\left(\hat{R}(v)\right)$$

The MSE of these measures is given by

$$\varepsilon(\vartheta) = \frac{1}{10\,000} \sum_{s=1}^{10\,000} \left| \frac{\hat{\vartheta} - \vartheta}{\vartheta} \right|_{G_s}; \ \vartheta = GD, APL, R, PM$$

Table I presents the MSE of the similarity measures.

	T_{GD}	T_{APL}	T_R	T_{DG}	$T_{R\varepsilon}$	$T_{DG\varepsilon}$	T_{GD}	T_{APL}	T_R	T_{DG}	$T_{R\varepsilon}$	$T_{DG\varepsilon}$	T_{GD}	T_{APL}	T_R	T_{DG}	$T_{R\varepsilon}$	$T_{DG\varepsilon}$
			1000						100						10			
VS	1,19	1,11	2,18	1,11	0,11	0,19	1,11	1,10	0,19	0,19	0,19	0,19	1,18	1,11	0,11	1,49	0,11	1,11
VSN	0,11	0,17	0,11	0,60	0,10	0,19	0,19	0,11	0,19	0,11	0,16	0,10	0,11	0,11	0,17	0,11	0,10	0,18
RSS-WR	0,11	0,19	0,11	0,18	0,10	0,81	0,18	0,11	0,11	0,19	0,41	0,11	0,49	0,19	0,19	0,18	0,79	0,11
RSS-WOR	0,19	0,18	0,80	0,19	0,19	0,11	0,61	0,61	0,69	0,49	0,61	0,19	0,61	0,71	0,18	0,61	0,48	0,10
BFS	0,18	0,11	0,11	0,11	0,61	0,11	0,11	0,18	0,11	0,49	0,10	0,19	0,10	0,61	0,10	0,61	0,48	0,10
RFS	0,11	0,11	0,18	0,11	0,11	0,61	0,11	0,18	0,11	0,61	0,49	0,11	0,19	0,10	0,18	0,11	0,61	0,11
SBS	0,19	0,18	0,11	0,10	0,11	0,12	0,19	0,11	0,11	0,11	0,41	0,19	0,19	0,48	0,10	0,11	0,61	0,11
FFS	0,10	0,41	0,11	0,19	0,19	0,11	0,91	0,11	0,11	0,19	0,18	0,11	0,40	0,41	0,11	0,18	0,41	0,19

Table I: Relative Mean Absolute Error of path measures

The MSE´s values are not too different when analysing the role of B in each measure. RSS-WR and RSS-WOR are the best sampling design in terms of the accuracy. VS and VSN have the worst results.

The evaluation of the behaviour of the normal approximation is given in the next table.

	T_{GD}	T_{APL}	T_R	T_{DG}	$T_{R\varepsilon}$	$T_{DG\varepsilon}$	T_{GD}	T_{APL}	T_R	T_{DG}	$T_{R\varepsilon}$	$T_{DG\varepsilon}$	T_{GD}	T_{APL}	T_R	T_{DG}	$T_{R\varepsilon}$	$T_{DG\varepsilon}$
			3000						300						50			
VS	0,63	0,65	0,68	0,61	0,54	0,72	0,55	0,50	0,53	0,59	0,49	0,69	0,48	0,51	0,51	0,54	0,46	0,67
VSN	0,66	0,76	0,81	0,60	0,50	0,72	0,53	0,57	0,63	0,55	0,44	0,70	0,51	0,55	0,56	0,57	0,40	0,68
RSS-WR	0,54	0,69	0,61	0,58	0,55	0,86	0,48	0,58	0,61	0,43	0,47	0,77	0,42	0,52	0,59	0,48	0,45	0,74
RSS-WOR	0,59	0,58	0,80	0,59	0,63	0,61	0,44	0,45	0,73	0,49	0,45	0,53	0,41	0,46	0,68	0,46	0,48	0,50
BFS	0,68	0,75	0,65	0,66	0,66	0,66	0,54	0,68	0,54	0,43	0,50	0,62	0,50	0,64	0,50	0,44	0,48	0,60
RFS	0,61	0,61	0,78	0,61	0,56	0,71	0,55	0,58	0,66	0,57	0,43	0,67	0,49	0,60	0,68	0,51	0,41	0,65
SBS	0,73	0,68	0,67	0,60	0,50	0,70	0,69	0,54	0,54	0,54	0,41	0,69	0,62	0,48	0,50	0,56	0,44	0,66
FFS	0,40	0,57	0,61	0,62	0,63	0,76	0,34	0,44	0,57	0,52	0,48	0,71	0,30	0,41	0,55	0,48	0,44	0,72

Table II. Estimated $\gamma\ for\ 1-\alpha = 0{,}95$ in normal tests on path measures

As noted, the above tables suggests that the approximation of the test statistics to normal distributions is not satisfactory in any case. The differences between the tests when *B=300* and *B=50* are very small.

13. Measures of similarity

The existence of similarities may be evaluated by means of ideas coming from statistical taxonomy and clustering.

$$\widehat{\overline{MS}} = \frac{1}{|G_s|}\sum_{v\in G_s} MS\,(v|G_s); MS = CloC, BC, r, CC$$

$$\overline{MS} = \frac{1}{|G|}\sum_{v\in G} MS\,(v|G); MS = CloC, BC, r, CC$$

which are estimated using the naive estimators

$$\widehat{\overline{MS}} = \frac{1}{|G_s|}\sum_{v\in G_s} MS\,(v|G_s); MS = CloC, BC, r, CC$$

$$T_{MS} = \frac{\widehat{\overline{MS}} - \overline{MS}}{\sqrt{\hat{\varepsilon}\left(\widehat{\overline{MS}}\right)}}$$

Another important measure is the Average Clustering Coefficient. The proposed test statistic is

$$T_G = \frac{\hat{C}(G_s) - C(G)}{\sqrt{\hat{V}(\hat{C}(G_s))}}$$

The MSE of the measures are calculated using

$$\varepsilon(\vartheta) = \frac{1}{10\,000}\sum_{s=1}^{10\,000}\left|\frac{\hat{\vartheta}-\vartheta}{\vartheta}\right|_{G_s};\ \vartheta = CloC, BC, r, CC\ ,\ \mathrm{G}$$

	T_{CloC}	T_{BC}	T_r	T_{CC}	T_G	T_{CloC}	T_{BC}	T_r	T_{CC}	T_G	T_{CloC}	T_{BC}	T_r	T_{CC}	T_G
			3000					300					50		
VS	1,65	1,66	1,61	1,66	1,61	1,66	1,61	1,66	1,61	1,61	1,66	1,68	1,65	1,66	1,61
VSN	1,66	1,65	1,61	1,65	1,61	1,65	1,61	1,51	1,66	1,65	1,61	1,66	1,61	1,66	1, 71
RSS-WR	0,56	0,65	0,66	0,86	0,75	0,65	0,96	0,86	0,65	0,66	0,61	0,69	0,65	0,61	0,68
RSS-WOR	0,44	0,32	0,61	0,66	0,66	0,66	0,91	0,96	0,61	0,65	0,65	0,63	0,61	0,66	0,61
BFS	1,35	1,65	1,66	1,61	1,61	1,63	1,63	1,61	1,67	1,86	1,56	1,65	1,65	1,66	1,78
RFS	0,66	1,60	1,61	1,06	1,61	0,86	1,61	1,66	1,11	1,86	1,06	1,61	1,71	1,56	1,81
SBS	0,96	1,45	1,56	1,36	1,66	1,65	1,52	1,65	1,46	1,65	1,61	1,56	1,51	1,69	1,66
FFS	0,79	1,61	1,66	1,65	1,65	0,61	1,76	1,61	1,66	1,61	1,01	1,65	1,59	1,71	1,81

Table III: Relative Mean Absolute Error of Similarity Measures

The MSE values are similar for the different values of B in each measure. RSS-WR and RSS-WOR are the best sampling designs in terms of the accuracy. VS and VSN have the worst results.

	T_{CloC}	T_{BC}	T_r	T_{CC}	T_G	T_{CloC}	T_{BC}	T_r	T_{CC}	T_G	T_{CloC}	T_{BC}	T_r	T_{CC}	T_G
			3000					300					50		
VS	0,74	0,83	0,71	0,83	0,90	0,70	0,81	0,68	0,80	0,92	0,58	0,77	0,53	0,74	0,90
VSN	0,77	0,85	0,72	0,85	0,91	0,70	0,83	0,62	0,81	0,92	0,51	0,76	0,44	0,77	0,90
RSS-WR	0,73	0,85	0,67	0,83	0,95	0,68	0,80	0,60	0,81	0,96	0,60	0,71	0,52	0,72	0,91
RSS-WOR	0,85	0,89	0,62	0,63	0,93	0,77	0,82	0,58	0,57	0,92	0,62	0,74	0,53	0,44	0,94
BFS	0,85	0,86	0,79	0,82	0,90	0,71	0,81	0,69	0,71	0,91	0,63	0,77	0,57	0,59	0,93
RFS	0,88	0,78	0,71	0,79	0,91	0,75	0,73	0,70	0,73	0,89	0,62	0,72	0,61	0,68	0,91
SBS	0,90	0,74	0,67	0,65	0,97	0,80	0,76	0,62	0,55	0,87	0,77	0,70	0,58	0,50	0,94
FFS	0,91	0,82	0,87	0,69	0,95	0,89	0,87	0,74	0,59	0,92	0,69	0,81	0,66	0,43	0,91

Table IV: Estimated $\gamma \; for \; 1 - \alpha = 0{,}95$ in tests on Similarity Measures

As expected, T_G behaves as a standard normal variable. T_{CloC} and T_r have a non-normal distribution even for B=3000. Hence, the inferences based on accepting a CLT would be very bad. From the other measures, assuming normality is not a good decision.

Acknowledgments

The authors acknowledge the support of CITMA project PN223LH010-033.

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. (revised May, 2020).

Arnab R. (2017): Survey sampling theory and applications, Academic Press, N. York.

Ahmed, N., J. Neville, and R. R. Kompella (2011). Network sampling via edge-based node selection with graph induction. Technical Report CSD TR 11-016, Computer Science Department, Purdue University, 2011.

Ahmed, N. J. Neville, and R. Kompella. (2012): Network sampling designs for relational classification. In Sixth International AAAI Conference on Weblogs and Social Media, 2012.

Birnbaum, Z. and M. Sirken (1965). Design of sample surveys to estimate the prevalence of rare diseases: Three unbiased estimates. Vital and Health Statistics, PHS Publication No. 1000-Series 2, No. 11. National Center for Health Statistics, Washington, D. C.: U. S. Government Printing Office.

Birnbaum, Z. and Agarwal G. and D. Kempe (2008): Modularity-maximizing graph communities via mathematical programming. The European Physical Journal B-Condensed Matter and Complex Systems, 66(3):409–418.

Blagus, N., L. Subelj, G. Weiss, and M. Bajec (2015): Sampling promotes community structure in social and information networks. Physica A: Statistical Mechanics and its Applications, 432:206 – 215.

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

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

Chiericetti, F., A. Dasgupta, R. Kumar, S. Lattanzi, and T. Sarl´os (2016): "On sampling nodes in a network," in Proc. 25th Int. Conf. WWW '16, 471–481.

Chin, A.W.H.; Chu, J.T.S.; Perera, M.R.A.; Hui, K.P.Y.; Yen, H.-L.; Chan, M.C.W.; Paris, M.; Poon, L.L.M. (2020): Stability of SARS-CoV-2 in different environmental conditions. Lancet Mircobe 2020, 20.

Frank, O. (1977). Estimation of graph totals. Scandinavian Journal of Statistics 4, 81–89.

Frank, O. (1980a). Estimation of the number of vertices of different degrees in a graph. Journal of Statistical Planning and Inference 4, 45–50.

Frank, O. (1980b). Sampling and inference in a population graph. International Statistical, Review 48, 33–41.

Frank, O. (2011). Survey sampling in networks. The SAGE Handbook of Social Network Analysis.

Frank, O. and T. Snijders (1994). Estimating the size of hidden populations using snowball sampling. Journal of Official Statistics 10, 53–67.

Ferguson M.N. et al. (2020): Impact of non-pharmaceutical interventions (NPIs) to reduce COVID19 mortality and healthcare demand. https://www.imperial.ac.uk/media/imperial-college/\medicine/ sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf. (accessed May 2020).

Friggeri, A., G. Chelius, and E. Fleury (2011): Triangles to capture social cohesion. In Privacy, security, risk and trust (passat), 2011 IEEE Third International Conference on Social Computing (Social.com), 258–265.

Fuller W. A. (2009): Sampling Statistics, Wiley, Chichester.

Goldenberg, A., Zheng, A.X., Fienberg, S.E., Airoldi, E.M. (2010): A survey of statistical network models. Found. Trends Mach. Learn. 2, 129–233.

Goodman, L. (1961). Snowball sampling. The Annals of Mathematical Statistics 32, 148–170.

Harvey N. (2011): Graph sparsifiers: A survey. presentation slides, 2011.

Kucharski A., Russell T., Diamond C., Liu Y. (2020): Analysis and projections of transmission dynamics of nCoV in Wuhan. CMMID repository, 2.

Kurant, M., A. Markopoulou, P. Thiran (2010): On the Bias of Breadth First Search (BFS) and of Other Graph Sampling Techniques," International Teletraffic Congress, 2010.14].

Kurant, A. Markopoulou, and P. Thiran (2011): Towards unbiased bfs sampling. Selected Areas in Communications, IEEE Journal 29,1799–1809.

Lavallèe, P. (2007). Indirect Sampling. New York, USA: Springer Science and Business Media.

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

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

Leskovec J. and C. Faloutsos (2006): Sampling from large graphs. In Proceedings of the 12th ACM SIGKDD international conference on knowledge discovery and data mining, pages 631–636. ACM, 2006.

Lewis T. (2011): Network science: Theory and applications. Wiley, Chichester.

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.

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.

Moitra. A. (2009): Vertex sparsification and oblivious reductions. FOCS, 2009.

Müller, J., Kretzschmar, M. & Dietz, K. (2000): Contact tracing in stochastic and deterministic epidemic models. Math. Biosci. 164, 39–64.

Newman, M. E. J. (2010): Networks: An Introduction. University Press.

Pham, H. (2020): On Estimating the Number of Deaths Related to Covid-19, Mathematics 2020, 8, 655; doi:10.3390/math8050655.

Ribeiro B. and D. Towsley (2012): On the estimation accuracy of degree distributions from graph sampling. In Decision and Control (CDC), 2012 IEEE 51st Annual Conference 5240–5247.

Salganik M. J. and D. D. Heckathorn (2004): Sampling and estimation in hidden populations using respondent driven sampling. Sociological methodology, 34,193–240.

Sirken, M. (2005). Network sampling. Encyclopedia of Biostatistics. John Wiley and Sons, Ltd., N. York.

Stokes J. and S. Weber (2017): "On the number of star samples to find a vertex or edge with given degree in a graph," in 2017 51st Annual Conference on Information Sciences and Systems (CISS), 1–6.

Stokes J. and S. Weber (2017): "Star sampling with and without replacement," in Proceedings of the International Workshop on Mining and Learning with Graphs (MLG), August 2017.

Stumpf M. P. and C. Wiuf (2005). Sampling properties of random graphs: the degree distribution. Physical Review E, 72, 103-118.

Tang B., Bragazzi N.L., Li Q., Tang S., Xiao Y, Wu J. (2020): An updated estimation of the risk of transmission of the novel coronavirus (2019-nCov). Infectious Disease Modelling, 5, 248255.

Wang, T., Y. Chen, Z. Zhang, T. Xu, L. Jin, P. Hui, B. Deng, and X. Li (2011): Understanding graph sampling algorithms for social network analysis. In Distributed Computing Systems Workshops (ICDCSW), 2011 31st International Conference on,123–128.

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.

Yan B. and S. Gregory. (2013): Identifying communities and key vertices by reconstructing networks from samples. PLOS ONE, 8(4): e61006.

Zhang, L.-C. and M. Patone (2017). Graph sampling. Metron 75, 277–299. Bipartite incidence graph sampling.

Zhang L-Ch and M. Ǒguz-Alper (2020): Bipartite incidence graph sampling, arXiv: 2003.09467v1[stat.ME] 20 March 2020 (accessed May 15, 2020).

Zhigljavsky, A. et al. (2020): Generic probabilistic modelling and non-homogeneity issues for the UK epidemic of COVID-19. arXiv:2004.01991v1[stat.AP]4Apr2020. (accessed May 2020).

CHAPTER VI

A SUPERPOPULATION MODEL FOR IMPUTATION OF MISSING DATA IN THE STUDIES OF COVID-19

CARLOS N. BOUZA-HERRERA[1]
SIRA M. ALLENDE-ALONSO[1]
MIR SUBZAR[2]

Abstract

This paper presents some issues on statistical research for COVID-19 and some ideas on the imputation of missing data. The imputation procedures are based on particular superpopulation models. The behaviour of the proposals is illustrated by a numerical study.

Keywords: COVID-19, sampling, imputation of missing data, superpopulation models.

1. Introduction

Frequently, surveys are developed for finding rates in politics, economy, marketing etc. The need of evaluating the widespread of an epidemic poses similar problems. The same basic approach is using a random sample from the population. The question that intrigues a country population is how many people are or have been infected with the coronavirus. Health authorities are more concerned with which is the probability of contagion with an infected than with the probability of being infected by the virus.

[1] Facultad de Matemática y Computación, Universidad de La Habana, Cuba
[2] Division of Agricultural Statistics SKUAST-K, Shalimar, India

Commonly, studies on COVID-19 analyse symptomatic patients. This approach is the classic biased procedure discussed in statistical literature, as only the stratum of the symptomatic is providing data for the studies. The population of asymptomatic is not sampled, and this is an example of the classic non-sampling error of bad coverage. The real interest of epidemiologists is detecting people infected with coronavirus (who are a source of contagion) or people who has a large probability of becoming severe cases, in case of being infected by the contact with symptomatic people. Post-pandemic, one of the interest issues is evaluating recovered people and their contacts behaviour, for establishing the possible effects of being infected with the virus.

So, the problem of developing sample-based statistics does not rely on increasing the number of sampled people, but in developing an adequate tool for identifying who needs to be tested.

Statistical inference is based upon and bounded by a theory developed on the basis of a certain approach. Frequentist and Bayesian are the most popular approaches. Sampling theory considers that the uncertainty in the data is introduced by the sampling design. The statistician determines the sampling design and the needed inferential procedures are developed.

- The selection of the sample determines a set of individuals. They should provide information on variables of interest. Commonly, some of the measurements are not obtained by different causes. List-wise deletion is the default method implemented in most statistical software packages. But excluding some cases, the statistical analysis is doubtfully correct. A solution is to substitute the missing data using imputation methods. Imputation provides complete data for developing subsequent analysis. As the imputed data incorporate the needed information, it is expected that the analysis would be statistically efficient and coherent – see Schenker and Raghunathan (2007) for a discussion of these aspects.

Commonly, data are gathered from a finite population U. It may be referred to as being "generated by a stochastic model". Hence, we consider that it is a realisation from a superpopulation model M. The presentation of data-based statistics commonly relies on methods based on considering that they are coming from an infinite population. Further discussions do not picture that the data come from the enumeration of a finite population. Finite sampling theory deals with a well-defined finite population $U = \{u_1, \dots u_N\}$.

Each unit u_i is perfectly identifiable in advance by the researcher. The variables involved changes; the population is dynamic. In this context, the concept of superpopulation is needed to differentiate between a finite population and an infinite superpopulation.

In sampling from a finite population, we often may find a reasonable probability model ("superpopulation model") that characterises relations among variables of the units of the population. For example, a physician with experience has knowledge of how each patient will recover from a certain variable in his records.

We are going to use a superpopulation regression model, where a guessed regression coefficient is fixed, as an imputation mechanism to estimate totals. We hypothesise that the researcher may determine a plausible superpopulation model. This superpopulation is determined by fixing a value of the involved parameters. The values of them are to be predicted (guessed) in basis of the experience of the researcher. Hence, the method we are considering substitutes each missing value by a prediction, which depends on the parameter that is fixed by the researcher.

We analyse the behaviour of our proposal for determining a total under missing observations. The effect of the accuracy off the guessed parameter in the statistical analysis is determined in terms of biases and mean squared errors. We study them under the use of model inference, at first. Afterwards, the sampling design effect is calculated, when random samples are considered as a new source of uncertainty. Therefore, the design expectations of the model bias and variance are computed. Finally, regularities of the expectations under a particular non-response mechanism are established.

The next section presents some issues on statistical research of COVID-19. Section 3 is concerned with discussing some ideas on the philosophy of imputation procedures and superpopulation modelling. In the fourth section we present results of a study of Bouza et al. (to be published) on the prediction of the sample total under a MCMAR missing observation mechanism. Section 5 presents a model for sampling graphs. It allows the identification of contacts with infected people and to calculate the risk of developing severe complications if the virus is caught. Finally, a discussion on the behaviour of the predictions is developed using a numerical study.

2. Some issues on different COVID-19 published studies

Peymane-Cheng (2020) has developed a meta-analysis with 46,248 patients confirmed with COVID-19 by laboratories. The significant observed odds ratio of people with severe disease were:

Disease	Odds ratios	95% confidence interval
hypertension	2.36	1.46 to 3.83
respiratory disease	2.46	1.76 to 3.44
cardiovascular disease	3.42	1.88 to 6.22

Table 6.1. Odds ratios reported by Peymane-Cheng (2020)

Some other studies detected as risky conditions obesity and smoking. Men present complications more frequently than women. The differentiation of subsequent comorbidities by sex seems to be influenced by a different social behaviour. The relative importance of other underlying health and social conditions are confounding factors.

These studies have been mainly conducted among hospitalized patients, with high risk, who received full testing. Clearly this is because the top priorities of public health systems are developing tests in people in hospitals and in medical staff involved with the pandemic. That is, the studies use data provided by testing symptomatic people. Therefore, these findings are not generalizable to the population.

Due to the use of selective testing, we are not able to argue on the true extent of the contagions and, consequently, on the virus's infection rate. Being uninformed about the number of infected people, we are not confident in the statistics on case fatality rate (probability of death on acquiring the virus) as well as many other statistics.

As stated before, studies on COVID-19 analyse symptomatic patients. This approach is the classic biased procedure discussed in statistical literature, as only the stratum of the symptomatic is providing data for the studies. The population of asymptomatic people is not sampled, and this is an example of the classic non-sampling error of bad coverage. The real interest of epidemiologists is detecting people infected with coronavirus (who are a source of contagion) or people who have a large probability of becoming severe cases, in case of being infected by the contact with symptomatic

people. That is the case both during the development of the pandemic and in post-pandemic studies.

The problem of developing sample-based statistics does not rely on increasing the number of tests, but in developing an adequate tool for identifying who needs to be tested.

Testing symptomatic patients exemplifies the classic un-coverage non-sampling error. People tested for the coronavirus are not a representative sample for evaluating the status of the population. Therefore, the data does not provide a good basis for arguing on the rate of infection and case fatality rate for the population.

Researchers who want to detect who has coronavirus need to identify groups of people with higher risk rates of infection.

The reasons for this selective testing are completely understandable. When testing is a scarce resource, people with COVID-19 symptoms should be preferably tested. Time and health workers are both limited, and it is convenient using the resources with symptomatic people who go to hospitals and/or doctor's offices. Clearly, if you are using health facilities, you are more likely to be detected as symptomatic of having COVID-19. So, we may consider the detected infected people receiving proper treatments, but they may also be considered for tracing their contacts.

Public health policy will never recommend randomly picking people from across the whole population and testing them for determining the presence of coronavirus. On developing testing in a representative sample, it is possible to obtain data for developing statistics on different aspects of the infection. The problem to be solved, at first, is to determine how many people are needed to be randomly tested, for getting data that describe accurately the situation in a population. The mathematical tools for answering this question have been worked out since 1934. The number of individuals is probably smaller than expected, particularly if we consider the use of some specialised method. We may consider following up the contacts with people infected for identifying high-risk groups and evaluate the fraction of those positive for the coronavirus.

A national random sample will allow epidemiologists to learn much more on the extent of the epidemic. But the real interest should be identifying people who are infected, but not sick, for ending with the chain of

contagious. Such samples would also provide information with respect to geography, ethnicity, sex and other demographic variables. The COVID-19 infection rates and its case fatality rate may vary across different regions and demographic variables. Hence using such methods could illuminate hidden trends before the damage is done. A main point is that public health officials, using the derived statistics for high-risk groups or regions, can be able to enact targeted and nuanced policies to help. Most countries in Latin-America and India have similar climate conditions though the management of the COVID-19 pandemic was very different. Nevertheless, the modelling we have presented here may be used in both regions in further studies on the pandemic.

Public health officials have used random samples in other settings. The theory of random sampling has proved it works effectively in areas of polling (politics, marketing etc.) as well as in scientific research (design of experiments, etc.). The only thing public health officials must cope with is the high cost of testing the presence of the virus.

Section 2 introduces some issues, which characterises aspects of the research that motivate the use of sampling to cope with missing data. Section 3 is concerned with providing a framework for dealing with imputing the missing data, and how a Bayesian approach may provide the needed tools by using superpopulation models in this context. Section 4 develops concrete ideas on the procedures for predicting missing data. Finally, a numerical study based on considering risky groups is discussed.

3. Some COVID-19 issues

COVID-19 virus belongs to a large family of viruses causing different illness going from common colds to severe diseases. We may quote Middle East Respiratory Syndrome (MERS-CoV) and Severe Acute Respiratory Syndrome (SARS-CoV) among severe diseases. COVID-19 is a new pandemic and the virus spreads from close contacts among people. Respiratory droplets released via coughs/sneezes is the main source of contagion. Nevertheless, it may be possible to be infected by touching a surface or object having the virus on it and touching the mouth, nose, and/or eyes afterwards. The initial symptoms range from fever, cough and breath shortness to more severe ones. They may appear from two to 14 days after exposure to the virus. The complications leading to death are commonly pneumonia in both lungs, particularly in risky-groups such as elderly people, immune-compromised/immune-suppressed individuals as well as in

people having some underlying health conditions (cardiologic, cancer, diabetes, etc.). External Risks are dependent on exposure. For example, workers involved in health care for COVID-19 patients and others in close contact with such patients have a higher risk of infection.

One of the challenging issues to deal with is that some infected people are asymptomatic (have experienced no symptoms at all) but they may spread the virus. Some patients may also be a source of contagion after having overcome the disease. Detecting them is one of the main interests of epidemiologists, both when the pandemic is still under development, for stopping the increase of the contagion, as well as in the post-pandemic management, for avoiding outbreaks.

Epidemiologists must develop inquires and face similar problems, as in any inquiry some of the people visited may not be evaluated. Elicitating information for non-respondents is needed for developing appropriate statistics. Next section introduces procedures for imputing missing data, supported by guessing some parameters of a superpopulation model. The model is selected by the researchers from a wide range of them.

4. Some issues on imputation and superpopulation procedures

In common statistical practice we deal with a finite population U of N units: $U = \{u_1, \dots, u_N\}$. Each unit j provides a pair of values X_j, Y_j . The X-variable is known or obtainable for any j, while for some units Y may be missing. Consider X as information on the people provided by medical records and Y the variable of interest as the result of the test.

The total population

$$T = \sum_{i \in U} Y_i$$

is of interest for the epidemiologist, but a census may not be developed. It is usual that the researcher obtains a sample s for analysing the behaviour of Y. Different approaches to point estimation may be adopted in the presence of non-response. Some methods just ignore the non-response. In the case of non-response unit, this will usually involve treating the set of responding units as if it were the selected sample.

Having non-responses generates three possible decisions:

1. Use only the available data
2. Select a subsample among the missing observations
3. Impute the missing values of Y.

The first decision is very risky, as the non-respondents may have a completely different behaviour than the respondents. In such cases, subsampling is the best solution from the statistical point of view, but it is costly. Having an adequate imputation method may solve the problem. Using the information on X and of the obtained data, the statistician may consider that having the total of the data in the sample allows characterizing the problem under study. In many applications this is the main objective of the inquiry. We will consider that the researcher in principle does not need computing the total of the variably Y in the population. Hence, we compute it in the sample

$$t = \sum_{j=1}^{n} Y_j = \sum_{j \in s} Y_j$$

and its knowledge allows to evaluate the phenomena behaviour. For example, the physician is evaluating the result of a medicament (as a vaccine) and the interest is evaluating if it provides the expected improvement in the patients. A function g (.) is evaluated. The number of evaluated people may play an important role in the evaluation of it, as some methods are sensitive to the lack of units. That is the case in many medical researches. The investigation must be based on a certain minimum number of observed units for being credible or for having sufficiently large degrees of freedom.

Missing data is a problem, which arises in many real-world applications of statistics. Imputation of the missing values is a way to deal with data set. Rubin (1976) fixed the existence of three possible types of missingness mechanism:

- Missing Completely at Random (MCAR)
- Missing at Random (MAR)
- Missing Not at Random (MNAR).

MCAR and MAR are in the class of ignorable missingness mechanism, but MNAR is a non-ignorable type of mechanism. Though MCAR assumption

is generally difficult to meet, in fact in some studies, where the sampling units are under control, as in a laboratory or in particular populations with patients in an experiment with a new drug, an experienced researcher may support that there is no statistically significant difference between incomplete and complete cases. He/she considers that in such a study missingness is due to a chance mechanism.

Commercial softwares include some imputation methods in their library, as for example: ICE, Imputation with Chained Equations (Stata), SAS IVEware: Imputation and Variance Estimation Software, R Packages (MICE, Amelia, missForest, Hmisc, mi) – see Raghunathan et al. (2016), Waljee et al. (2013) and Rickert (2016).

The evaluation of the sample determines whether a unit provides information or not. This collection of random variables is called a superpopulation. Considering that the behaviour of the random variables is described by a certain probability structure, this structure is often stated in terms of the so-called superpopulation model. Deming and Stephan (1941) introduced the term, but Cochran, in 1939, used this approach at first – see Cochran (1946). They agree in considering that the finite population under study was drawn from a larger universe. In such cases the parameters of the superpopulation have a statistical meaning, as different sets of N subjects will arise from the realisation for the superpopulation. Särndal (1992) used this concept of superpopulation to consider that it is an abstract representation of "a broader entity from which the population values are generated". In this context, the superpopulation represents a causal system. Then, the potentially observed random variables and the missing ones are described by the superpopulation model.

The researcher should be able to fix such probabilistic structure, the involved assumptions, and a consensus can be reached as to constitute the 'best' guess on the model to be assumed.

5. Predicting under missing data

Consider that we have a sample of n individuals selected from a population of size N, then the total sample of the variable of interest Y is given by

$$t = \sum_{i=1}^{n} y_i$$

Assume that an auxiliary variable X is known for all the individuals in the population. This variable is related with Y by means of a superpopulation model. That is commonly the case in medical research, where X appears in the files of the patients. We will consider the superpopulation model

$$M\colon\ Y_i = BX_i + e_i, E(e_i e_j) = \begin{cases} \sigma^2\ if\ i = j \\ 0\ otherwise \end{cases}$$

Commonly, the information of some sample units is missing, and it is not computable. Some notation is needed. Take

$s_1 = \{i \in s | y_i\ is\ obtained\}$, $s_2 = \{i \in s | y_i\ is\ missing\}$, $n_j = |s_j| > 0, j = 1,2$

$$\bar{z}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} y_i\ , j = 1,2; z = x, y$$

We only can compute the mean of Y for the individuals in s_1, but the mean of X may be obtained for the whole sample $s = s_1 + s_2$. The total sample t may be rewritten as

$$t = \sum_{i=1}^{n} y_i = n_1\bar{y}_1 + n_2\bar{y}_2 = t_1 + t_2$$

Using the information obtained and the superpopulation model M, we may impute the missing values of Y by using the predictor $\hat{y}_i = Bx_i$. Generally, B is unknown. Consider that B may be approximated using available information for guessing an appropriate value B_0. For example, in a longitudinal study, the physician may fix a "guessed vale" B_0. The physician may be considering that the value of B obtained in the previous evaluation of the patients should, say B_a, is incremented in such a way that $B_0 = \gamma B_a$.

From these facts, and using the superpopulation M, what is possible is to use $\hat{y}_{i0} = B_0 x_i$ and compute.

$$t_{02} = n_2\bar{y}_{20} = \sum_{i=1}^{n_2} B_0 x_i$$

The respondents may be used for having an idea of the value of the residuals in the predictions made by using the guessed parameter B_0. The "guessed" residual is $e_{0i} = y_i - B_0 x_i$. They may be computed only if $i \in s_1$. The e_{0i}`s may be used for centering the predictor t_{02}. Our proposal is using as predictor of t

$$t^* = n_1 \bar{y}_1 + n_2 \left[\frac{1}{n_1} \left(\sum_{i=1}^{n_1} y_i - B_0 x_i \right) + \frac{1}{n_2} \sum_{i=1}^{n_2} B_0 x_i \right]$$

using this predictor of the model bias – see Bouza et l (to be published)

$$E_M(t^* - t) = \delta \left[\frac{n_2}{n_1} \sum_{i=1}^{n_1} x_i - \sum_{i=1}^{n_2} x_i \right], \delta = B - B_0$$

Hence the model bias is $B_M = n_2 \delta [\bar{x}_1 - \bar{x}_2]$ and the proposed predictor is approximately unbiased if:

1. $B \cong B_0$, say that the guessed value of the parameter is close to the true B, $or\ if\ \bar{x}_1 \cong \bar{x}_2$, say that the strata of non-respondents and of respondents have a similar value of the auxiliary variable.
2. $E(E_M(t^* - t)|s) = n_2 \delta (\mu_{X_1} - \mu_2)$

The first condition establishes that if the respondent and non-respondent strata have the same mean of the auxiliary variable, the expected bias is zero. The second condition may be checked once the sample is evaluated. Clearly n_2 is distributed according to the binomial $B(n, W_2)$, then

$$E[E_d(E_M(t^* - t)|s)] = nW_2 \delta (\mu_{X_1} - \mu_2) = Bias(t^*)$$

The model ´s mean squared error, MMSE, is

$$E_M(t^* - t)^2 = n_2^2 E_M \left\{ \left[\frac{1}{n_1} \left(\sum_{i=1}^{n_1} \delta x_i + e_i \right) - \frac{\delta}{n_2} \sum_{i=1}^{n_2} x_i \right]^2 - 2 \left[\frac{1}{n_1} \left(\sum_{i=1}^{n_1} \delta x_i + e_i \right) - \frac{\delta}{n_2} \sum_{i=1}^{n_2} x_i \right] \bar{e}_2 + \bar{e}_2^2 \right\}$$

denoting

$$H = \left[\left(\frac{1}{n_1}\right)^2 \left(\sum_{i=1}^{n_1} \delta x_i + e_i\right)^2 + \left(\frac{\delta}{n_2} \sum_{i=1}^{n_2} x_i\right)^2 - \frac{2}{n_1 n_2} \left(\sum_{i=1}^{n_1} \delta x_i + e_i\right) \sum_{i=1}^{n_2} \delta x_i \right] = A + B - C$$

$$A = \left(\frac{1}{n_1}\right)^2 \left[\delta^2 \{ \sum_{i=1}^{n_1} x_i^2 + \sum_{i \neq j=1} x_i x_j \} + \left(\sum_{i=1}^{n_1} e_i^2 + \sum_{\neq j=1} e_i e_j \right) + 2 \sum_{i=1}^{n_1} \delta x_i \sum_{i=1}^{n_1} e_i \right]$$

where

$$E_M(A) = \left(\frac{\delta}{n_1}\right)^2 [\sum_{i=1}^{n_1} x_i^2 + \sum_{i \neq j=1} x_i x_j] + \frac{\sigma^2}{n_1}$$

$$E_M(B) = \left(\frac{\delta}{n_2} \sum_{i=1}^{n_2} x_i\right)^2$$

$$C = \frac{2}{n_1 n_2} \left(\delta^2 \sum_{i=1}^{n_1} x_i \sum_{i=1}^{n_2} x_i + \delta \sum_{i=1}^{n_2} x_i \sum_{i=1}^{n_1} e_i \right)$$

$$E_M(C) = \frac{2\delta^2}{n_1 n_2} \left(\sum_{i=1}^{n_1} x_i \sum_{j=1}^{n_2} x_j \right)$$

Then

$$E_M(H) = \delta^2[(\bar{x}_1 - \bar{x}_2)^2] + \frac{\sigma^2}{n_1}$$

As a result, the MMSE is

$$E_M(t^* - t)^2 = n_2^2\delta^2[(\bar{x}_1 - \bar{x}_2)^2] + \frac{n_2 n\sigma^2}{n_1}$$

Note that having a guessed parameter close to B, or the means of the two samples being similar, the MMSE depends only on the ratio of the subsample sizes and on the model variance. For a relatively large number of respondents, $n_1 \gg n_2$, the model error will be smaller.

The Decision Maker may decide to use only the Model criteria for predicting t. In such case, if $\bar{x}_1 - \bar{x}_2 \cong 0$, the error is only a function of the last term.

$$U = U_1 \cup U_2;\ U_1 = \{i \in U | a\ response\ is\ obtained\ \},$$
$$U_2 = \{i \in U | a\ response\ is\ not\ obtained\},$$

$$N_j = \left|U_j\right| > 0, j = 1,2; N = N_1 + N_2\ .$$

Then we have that $W_j = \frac{N_j}{N}, j = 1,2$, are the response and non-response probabilities.

Note that for the derivations, the sample does not need to be random.

Consider that the sample was selected from U using as sampling design a simple random sampling with replacement.

$$E_d[E_M(t^* - t)|s] = \delta\left(\mu_{X_1} - \mu_{X_2}\right)$$

$$E_d[E_M(t^* - t)^2|s] = n_2^2\delta^2\left[\left(\mu_{X_1} - \mu_{X_2}\right)^2 + Var(\bar{x}_1 - \bar{x}_2)\right] + \frac{n_2 n\sigma^2}{n_1}$$

If we use SRSWR:

$$E_d[E_M(t^* - t)^2|s] = n_2^2\delta^2\left[\left(\mu_{X_1} - \mu_{X_2}\right)^2 + \frac{\sigma_{X_1}^2}{n_1} + \frac{\sigma_{X_2}^2}{n_2}\right] + \frac{n_2 n\sigma^2}{n_1}$$

From the derived results, we recommend using:

$$\bar{y}^* = \frac{t^*}{n} = w_1\bar{y}_1 + w_2\left[\frac{1}{n_1}\left(\sum\nolimits_{i=1}^{n_1}(y_i - B_0x_i\right) + \frac{1}{n_2}\sum_{i=1}^{n_2}B_0x_i\right];\ w_j = \frac{n_j}{n}, j = 1,2$$

It mimics Hansen-Hurvitz (1946) proposal, when subsampling the non-respondent stratum.

$EE_d[E_M(t^* - t)^2|s]$ is obtained approximately by using the approximations:

$$E\left(\frac{n_2^t}{n_1}\right) \cong \frac{E(n_2^t)}{E(n_1)} + V(n_1)\left(\frac{E(n_2^t)}{\left(E(n_1)\right)^3}\right) - 2\frac{Cov\,(n_2^t, n_2)}{\left(E(n_1)\right)^2} = \nu(n)$$

Then substituting the expectations, variance and covariance, we have, as an approximation to the overall MSE, for n large,

$$E\left(MSE(\bar{y}^*)\right) \to_n \delta^2\left(\frac{W_2^3}{nW_1}\sigma_{X_1}^2 + \frac{W_2\sigma_{X_2}^2}{n} + \right) + \delta^2\frac{W_2^3}{nW_1}\left(\mu_{X_1} - \mu_{X_2}\right)^2 + \frac{W_2}{nW_1}\sigma^2$$

When $\delta \cong 0$, the first two terms are negligible. If $\mu_{X_1} - \mu_{X_2} \cong 0$, we will have a small value of the second term of the error. If both relations hold, the EMSE only depends on the variance of the model error and on the relative size of the non-response stratum.

6. A numerical study

Physicians may consider the risk of individuals being positive to COVID-19 in terms of a function φ of the factors determined by evaluating risks

associated to age, chronic diseases and sex, say $\varphi(a, c, s)$. Individuals are stratified in terms of these.

According to age, we have 7 strata:

$$\begin{aligned} A(0) = babies, A(1) = children, A(2) = adolescents, A(3) \\ = people\ aged\ beween\ 19\ and\ 30\ years, A(4) \\ = people\ aged\ between\ 31\ and\ 60, A(5) \\ = people\ aged\ beween\ 61\ and\ 70\ years, A(6) \\ = older\ than\ 70\ years \end{aligned}$$

The health status determines other 7 strata:

$$\begin{aligned} Chronic\ diseases: C(0) = no\ disease, C(1) = being\ diabetic, C(2) \\ = being\ allergic, C(3) = being\ hypertense, C(4) \\ = being\ cardiopatic, C(5) \\ = having\ cancer\ disorders, C(6) = other\ diseases \end{aligned}$$

Sex determines two strata:

$$S(1) = man, S(2) = woman$$

Then each individual u_i belongs to a stratum:

$$U_{a,c,s} = A_a \cap C_c \cap S_s, a = 0,1, \dots ,6; c = 0,1, \dots ,6; s = 1,2$$

The rules implemented from a medical protocol permits to fix a risk:

$$\varphi_i(a, c, s) \in [0,1]$$

An interview or the use of the file u_i and

$$D(u_i; a; c; s) = \begin{cases} 1\ \ if\ u_i \in U_{a,c,s} \\ 0\ otherwise \end{cases}$$

The risk is measured by using a function of these factors and a coefficient of the current contacts with infected people G_i.

The simplest way of measuring risk in a person is given by:

$$R(u_i) = \frac{\sum_{s=1}^{2} \quad \sum_{a=0}^{6} \quad \sum_{c=0}^{6} D(u_i; a; c; s)\, \varphi_i(a, c, s) + G_i}{15}$$

The existing information is not 100% reliable. Individuals in a network, where at least one infected person is included, are to be visited and his status evaluated by a physician who tunes $D(u_i; a; c; s)$ substituting it by

$$W(u_i; a; c; s) = \begin{cases} a\ value\ in\ (0,1)\ \ if\ \ u_i \in U_{a,c,s} \\ 0\ otherwise \end{cases}$$

This value weights the importance of the present factors in terms of tests or subjective criteria of the visitor. For example, if $u_i \in A(j)$, the existing information allows establishing, during the visit, not only the existence or not, but how severe is the real, and it tunes its importance through $W(u_i; a; c; s)$. For example, if a person is classified in C(4) $\varphi_i(a, c, s)$, equal the importance of an angina attach of a person with another two surgical operations. The risk of u_i computed after the visit of the physician is:

$$Y(u_i) = \frac{\sum_{f=1}^{F} \quad \sum_{t=1}^{T} W(u_i; a; c; s))\varphi_i(a, c, s) + G_i}{15},$$

The protocols establish that u_i is to be evaluated as risky if $Y(u_i) > 0{,}9$ and then is to be controlled.

A person who is not interviewed generates a missing data. In our interviews with specialists, they considered that the superpopulation model provides an adequate tool for predicting the risk. The specialists considered that the parameter must be fixed within the strata. Then the model is by:

$$Y(u_i) = B_{a,c,s} R(u_i) + \varepsilon_i$$

Statistics from 7 specialized hospitals were obtained and the risk of 2,056 controlled people was measured. Using the information provided by them, it was used for fixing the value of B_{0k} and the missing observations are predicted by:

$$\hat{Y}(u_i) = B_{0a,c,s} R(u_i)$$

The missing observations are due to the failure of evaluating $Y(u_i)$ when visiting the patient. In our research, they were interviewed afterwards and $Y(u_i)$ was computed.

Take the parameter:

$$\bar{Y}(U_{a,c,s}) = \frac{1}{\|U_{a,c,s}\|}\sum_{j\in U_{a,c}s} Y\ (u_j)$$

A question is how good the accuracy of $\hat{Y}(u_i)\ is$. Say, if for any patient $\Delta(u_i) = \hat{Y}(u_i) - Y(u_i) \cong 0$. We evaluate the overall accuracy by computing:

$$\bar{\delta}\left(\bar{Y}(U_{a,c,s})\right) = \frac{1}{\|U_{a,c,s}\|}\sum_{j\in U_{a,c,s}} \left|\frac{\Delta(u_j)}{\bar{Y}(U_{a,c,s})}\right|$$

Some entries in the tables present a zero as entry because no observations were missing, or no case was observed. The best results for predicting within the A(j)´s are in red and in blue the best for the C(t)´s. The case in which the prediction was the best, both for the stratification by age and by chronic disease, is given in brown**.** The physicians considered that if:

$$\bar{\delta}\left(\bar{Y}(U_{a,c,s})\right) \leq 0{,}25$$

the prediction was acceptable.

The results of the research for men are in table 1. The babies had no missing data or observations for C (1), C (2), C (83) and C (5). The predictions for children exhibit the worst results. The oldest men (more than 70) behaviour was accurately predicted and those in C (5) followed them. Note that only 6,12% of the results would not be acceptable in terms of $\bar{\delta}\left(\bar{Y}(U_{a,c,s})\right)$.

Age Group	0	1	2	3	4	5	6	
0	21,4	0,0	0,0	0,0	0,0	0,0	23,5	
1	29,0	19,2	26,1	27,4	12,1	13,9	13,5	
2	19,4	13,5	19,2	23,4	9,3	8,0	12,5	
3	22,6	15,1	23,5	23,8	11,1	23,0	20,0	
4	21,3	8,9	26,4	26,6	5,3	11,8	14,2	
5	18,6	8,7	15,2	24,2	14,0	11,6	11,4	
6	18,1	6,1	15,2	21,1	15,7	20,7	8,3	

Table 6.2: Results of $100\bar{\delta}\left(\bar{Y}(U_{a,c,s})\right)$ in the analyzed men

Table 6.2 presents a similar analysis for women. Women aged more than 70 are expected to exhibit the more accurate results if they have no disease, not being diabetic, cardiopathic and not having psychological disorders. Females had the best results in the case of having no disease, not being allergic or presenting cancer disorders. Predictions for women having no disease exhibit the best performance. Note that the accuracy of the prediction in women is considerably better than with men. The prediction for men who were not allergic exhibits the best results. Note that, for the women, all the results are acceptable in terms of $\bar{\delta}\left(\bar{Y}\left(U_{a,c,s}\right)\right)$.

Age Group	0	1	2	3	4	5	6
0	7,2	10,1	11,2	11,3	0,0	0,0	15,0
1	9,9	10,0	6,1	0,0	0,0	17,0	17,5
2	13,9	10,5	13,2	16,4	14,1	25,9	22,1
3	12,5	13,0	9,7	15,7	27,1	16,5	20,6
4	11,5	9,9	17,2	14,5	12,8	23,7	11,1
5	10,8	13,8	10,6	19,6	18,0	11,3	20,8
6	9,7	9,7	13,7	12,5	10,7	11,1	11,5

Table 6.3: Results of $100\bar{\delta}\left(\bar{Y}\left(U_{a,c,s}\right)\right)$ in the analyzed women

Acknowledgements

The research producing this paper was supported partially by the projects cu2018JOt0044104. (A Cuban-Flemish Training and Research Program in Data Science and Big Data Analysis) and PN223LH010-005 (Desarrollo de nuevos modelos y métodos matemáticos para la toma de decisiones) .

Bibliography

Andridge, R. R. (2009): **Statistical methods for missing data in complex sample surveys**. Phd thesis, The University of Michigan.

Bouza-Herrera, C. N., C. Viada and G. K. Vishwakarma (2020): Studying the total under missingness by guessing the value of a superpopulation model for imputation. **Revista Investigacion Operacional,** Forthcoming 62d05-6-20-03.

bouza-Herrera, C. N., Allende-Alonso, S. M., G. K. Vishwakarma and N. Singh (2019): Estimation of optimum sample size allocation: An illustration with body mass index for evaluating the effect of a dietetic

supplement. **International Journal of Biomathematics** 12, 108-120.

Dorfman, A. H. and R. Valliant (2005): Superpopulation Models in Survey Sampling. **Encyclopedia of Biostatistics,** 1 Book Editor(s): Peter Armitage and Theodore Colton.

FOX, R, (2016): **Applied regression analysis and generalized linear models,** (3rd ed.), Dage Publications Inc, Thousand Oaks, CA.

GIRA, A. A. (2015): Estimation of Population Mean with a New Imputation Method. **Applied Mathematical Sciences**, 9, 1663 - 1672.

Jordan, R. E., P. Adab and K. K. Cheng (2020): Covid-19: risk factors for severe disease and death BMJ 2020*;* 368 doi: https://doi.org/10.1136/bmj.m1198 (Downloaded 7 June, 2020).

Little, R. J. A. and Rubin, D. B. (2002), Statistical Analysis with Missing Data, Wiley: New York, 2nd ed. Kabacoff, R. I. (2011) R in Action: Data analysis and graphics with R, In "**Advanced methods for missing data"**, Editor Manning, M. (chapter 15, 352-372).

Pham, H. (2020): On Estimating the Number of Deaths Related to Covid-19. **Mathematics**, 8, 655; doi:10.3390/math8050655.

Raghunathan, T., P. Solenberger, P. Berglund and J. van Hoewyk (2016): **IVEware: Imputation and Variance Estimation Software** (Version 0.3). Survey Research Centre, Institute for Social Research University of Michigan Ann Arbor, Michigan.

Rickert, J. (2016): Missing Values, **Data Science and R**.

Sarndal, C., Swensson, B. and Wretman, J. (1992): **Model Assisted Survey Sampling**. Springer-Verlag, New York.

Royston, P. (2004): Multiple imputation of missing values. **The Stata Journal**, 4, 227- 241.

Schenker, N. and T. Raghunathan (2007): Combining information from multiple surveys to enhance estimation of measures of health. **Statist. Med**. 26, 1802–11.

Waljee, A.K., A. Mukherjee, A. G Singal, Y. Zhang, J. Warren, U. Balis, J. Marrero, J. Zhu and P. D. R. Higgins (2013): Comparison of imputation methods for missing laboratory data in medicine https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3733317/ (Last consulted April 10, 2019).

Zhang, K., X. Liu, J. Shen, Z. Li, Y. Sang, X. Wu,Y. Zha, W. Liang, C. Wang, K. Wang, L. Ye, M. Gao, Z. Zhou, L . Li, J. Wang, Z. Yang, H. Cai, J. Xu and G. Wang (2020): Clinically Applicable AI System for Accurate Diagnosis, Quantitative Measurements, and Prognosis of COVID-19 Pneumonia Using Computed Tomograph. https://www.sciencedirect.com/science/article/pii/S0092867420305511 *(*Downloaded 7 June, 2020).

CHAPTER VII

THE SOCIOECONOMIC AND EDUCATIONAL IMPACT OF THE COVID-19 PANDEMIC ON INDIGENOUS STUDENTS AT THE INTERCULTURAL UNIVERSITY OF TABASCO, MEXICO

JOSÉ FÉLIX GARCÍA RODRÍGUEZ[1]
JOSÉ RAMÓN CONTRERAS DE LA CRUZ[2]
GUADALUPE MORALES VALENZUELA[3]
LOURDES DEL C. PINEDA CELAYA[4]
IGNACIO CAAMAL CAUICH[5]

Summary

The COVID-19 pandemic has negatively impacted the well-being, the population health and the economy on a global scale, also transforming the forms of human coexistence. Education has been one of the activities most affected by the social distancing adopted as a fundamental health strategy, which has impacted the quality and access to education at all levels, with greater emphasis on the most vulnerable population groups. The objective of the research is to know the opinion of the students at the Intercultural University of the State of Tabasco (UIET) about the impact of the pandemic in the educational process in its distance mode, adopted as an emerging

[1] Universidad Juárez Autónoma de Tabasco;
[2] Universidad Intercultural del Estado de Tabasco;
[3] Universidad Intercultural de Tabasco;
[4] Universidad Autónoma de Guadalajara;
[5] Universidad Autónoma Chapingo.

educational strategy from the measures of healthy distance and confinement, during the February-June 2020 semester. The research approach is qualitative and hermeneutical, supported by the application of semi-structured interviews with 70 students from the seven degrees offered by the institution. This experience allowed us to interpret how students perceived this health crisis and how it has affected their academic and family life, as well as to create recommendations for the improvement of educational practice in eventual emerging contexts.

Introduction

In March 2020, Mexico health and educational authorities, as well as other countries in the world, decreed school and work centre closures to avoid at all costs the massive contagion by the new COVID-19 disease caused by the coronavirus SARS-CoV2, whose high dissemination threatened to collapse hospital services. Only essential services such as health, food, transportation, water, electricity, construction, sanitation, cleaning, civil registration, among others, continued operating. The United Nations Educational, Scientific and Cultural Organization (UNESCO) reported that, as of April that year, 191 countries had closed their schools and 93% of their enrolment was no longer attending their schools (Castillo 2020). This phenomenon is unlike anything experienced in recent times. Activities around the world have stopped, and this health crisis is predicted to have consequences on the economy, food production and security in a period of between two to five years (ECLAC 2020).

The National Association of Universities and Institutions of Higher Education (ANUIES), which groups together 196 public and private universities in Mexico, took some measures from the beginning of the contingency; for example, it issued recommendations to face the pandemic and to return to the new normal through the document "Towards the collective construction of the new normal in higher education" (ANUIES, 2020a) and provided useful tools for academic work on its institutional page (ANUIES 2020b). Since its foundation in 2005, the Intercultural University of the State of Tabasco (UIET) had not seen such a drastic change in the training process of university students as the one experienced with COVID-19.

From March until the end of the school year (June of the same year), school activities were remote. This sudden change will result in immediate and medium-term effects in the academic life of Mexican university students,

especially those from families with low socioeconomic levels, and even more to students from multicultural backgrounds. This study was conducted with students from the UIET at the end of the semester (February-June 2020) to understand some of the lockdown effects by COVID-19 in their academic life and their viewpoint in this regard, to improve educational practice. We will focus on three aspects: inequity and higher education, resilience as a response to the crisis, and changes in the university educational process.

Inequality and higher education

The COVID-19 pandemic has made inequality visible in our societies, especially in Latin American countries, as reported by various authors, such as Vijil (2020) in Nicaragua, Vivanco (2020) and Vallejo (2020) in Ecuador, Failache (2020) in Uruguay, Tzoc (2020) in Guatemala, and Dietz (2020), Mérida and Acuña (2020) Buendía (2020), Acosta (2020) and CONEVAL (2020) in Mexico. This inequality has materialised in the absence of the means to continue training at almost all educational levels; UNESCO and other international organizations agree that in the coming months, and possibly years, health, economic and social crises will intensify, and the inequality gap will increase (Renna 2020, Acosta 2020 and ILO 2020).

The National Council for the Evaluation of Social Policy (CONEVAL 2020) mentions that this health crisis will have unavoidable impacts in the short and medium-term in Mexico, in aspects such as economy, trade, employment, and population welfare. Inequality in Mexico is reflected in the 49% of its population living in poverty and with a population with chronic presence of diseases such as diabetes, malnutrition, obesity, hypertension, as well as the income dependence of more than half of Mexican families in informal employment; 24.7 million people (19.8%) with lack of access to water, essential services and overcrowding in their homes (CONEVAL 2018). These last factors complicate the implementation of measures to contain the epidemic and violate the human right to water and health, as they cannot comply with the basic hygiene measures of washing their hands with soap and water and home isolation from the positive and suspicious (Mérida and Acuña 2020). This inequality is also observed in access to education, especially higher education; Acosta (2020) reports that this country is one of the worst rated in accumulated educational backwardness in the OECD. For example, in basic education, 30% of the children who enter primary school do not complete secondary school; and it is the last place of 35 countries of the ODCE in coverage in

higher education, in the order of 39%. The primarily cause in both cases is the low socioeconomic level of the families of student's origin.

Mérida and Acuña (2020) warn about the risk of inequity with regards to accessing the basic level for educational programs in some states with high levels of marginalization. For example, the federal and its counterparts "Learn at home" program from state governments did not consider the scarce and non-existing connectivity to television, internet, or radio in some regions of the country, which can lead to low levels of learning, increase the educational gap, inequality and social injustice. Similarly, Dietz (2020) notes that inequality will increase by the pandemic, but even more in access to education in contexts of cultural diversity. For this reason, intercultural schools and universities located in these multicultural regions can play a vital role in helping to reverse these inequalities. Some experts point out that higher education will undergo drastic changes from face-to-face to online training, as well as some areas of economic development, such as online work, industry, and tourism (Vijil 2020), along with the reduction of internationalization and student mobility (Mérida and Acuña 2020).

Resilience as an answer to the crisis

The native peoples of the Americas, especially in Latin American countries, are being impacted by this health crisis on different scales. Hooker (2020) states that "... they have seen different spheres of their internal life disrupted by being forced to make rapid and drastic changes in the organization of their ways of life that have affected their spirituality, their respect, how to deal with illness, death, and mourning, and in general the balance and harmony of their environment" (p. 2). Recognising that, in these societies, new capacities are emerging related to solidarity, resilience, and emotional intelligence to learn from events and occurrences, however painful they may be, and always return to the earth to take care of it and coexist with it. Following this same author, it is pointed out that intercultural universities in Latin America are experiencing, along with local communities, new ways of facing the crisis, with an already complex environment and the difficulties that existed before the pandemic, which included: a decrease in funding, non-quality growth, inequity to access infrastructure and services concerning other institutions of higher education system.

Dietz (2020) mentions the strategy that indigenous communities used in Veracruz to respond differently to the risk of the pandemic. Rather than using the expression "Stay at home", as it was mandatory by the federal

government, they secured access to their community to prevent foreigners from entering, even their migrant relatives, in a clear message as a community: "stay in community" which, according to him, can also apply in the educational sector, community learning. The same author reminds us that the native American peoples have demonstrated for centuries that they are resilience experts, facing all types of catastrophes and pandemics as a unified community. The advice of their sages and elders, and their biocultural memory, it is not to be doubted that they will bring out their best to face this new pandemic.

On the other hand, Tzoc (2020) mentions that Mesoamerican indigenous peoples have a multidimensional view of illness and death, and they face the crisis differently. Firstly, they respect all living beings, even those that cause suffering and disease; they have the notion that everything moves within periods of order and chaos and that human beings must seek the restoring of this order and harmony with nature; then that this pandemic is an opportunity to value life in a community, within the family, to seek local answers, learn, reconcile with the earth and nature and give meaning to life. In the same sense, Vallejo (2020) has been following the development of the pandemic among the Amazonians in Ecuador and has found that they have local ways to strengthen their respiratory system and defences with plants from the jungle. Despite that, the risk of infection by the SARS-CoV2 virus is high since coming from a wild species; human beings do not have defences for it. Cases are increasing and, in the absence of state health services, youth migration between the city and the countryside, and the presence of third parties for extractive, mining, oil, and logging activities, they fear the loss of the elderly ones with their knowledge and historical memory.

The Global Youth Survey 2020 (OIT 2020) has found that the impact of COVID-19 and confinement have been systematic, profound, and disproportionate for young people between the ages of 18 and 29, and that it will bring sequelae over time. It documents the effects of the pandemic in four areas: employment and income, education and training, mental well-being and diminished rights, the latter including freedom of worship, access to information, civic participation, and access to decent housing.

At UIET, this pandemic has altered the study programs and the planning lessons, as the suspension of face-to-face sessions created unforeseen and unfamiliar situations. Teachers and students were challenged by time and space. This confinement began with lockdown and social distancing and

caused a rupture in the school dynamics and in the culture of the school environment. The educational task became more complex when attending university students and their children's school activities, both for university students and university teachers. Despite the difficulties, complexities and shortcomings, students have faced these scenarios with the support of their families and professors. The training developed during their stay at the university and their life history of coming from families used to the daily struggle for survival have allowed them to face these difficulties with relative success. The results will show the findings that make us support the above.

Changes in the university educational process

In the beginning, it was supposed to be a temporary interruption of school and productive activities extended to prolonged confinement. Although the ongoing distance education activities were a temporary measure until in-person attendance could be re-established, with the passing of the weeks it became an extended measure and most of the educational systems were in trouble, the whole educational system, from the basic to university level. It forced us to consider transformations in the in-person-virtual paradigm in the changing times. Acosta (2020) points out that the academic inactivity of students who do not have the means can lead to a 10% decrease in learning and affect the passage to successive grades, graduation and employability. It is one of the main concerns of higher education institutions that serve young people from families of low socioeconomic levels, including intercultural universities.

Cotino (2020) approaches it from the right to education. He suggests turning virtual education from a lifesaver, in the health crisis, to an opportunity for change that takes better advantage of digital resources; he points out that, although in-person education is necessary, and in many ways superior to online education, for now this last one will allow the continuation of the student´s learning in the world. This same author recommends considering the edula@ab decalogue of the Universidad Oberta de Catalunya (EDUL@AB 2020) for better use of digital tools in education.

Acosta (2020) points out that in the case of higher education, each university has responded differently to the pandemic according to its available resources. For example, the Universidad Nacional Autónoma de México (UNAM) y la Universidad Autónoma Metropolitana (UAM) have implemented emergent programs to mitigate the negative impact of the

pandemic. In the case of the UAM, the institution has provided its low socioeconomic students with tablets and cards for their use in the trimester and implemented the Emerging Remote Teaching Project (UAM, 2020).

On the other hand, Gazzo (2020) recommends stopping looking at the cyberspace of virtual education as a black box or device to upload information files that the student will download and make use of it. Instead, she proposes to develop concrete but deep content, collaborative spaces, and maintain teacher-student interaction as a common thread throughout the process. The same author claims the millennials' role, who just a few weeks ago were repressed, punished, and singled out for depending on electronic devices, just in this sudden social shift pushed from biology by a tiny organism, that their skills will be of great use.

As expressed by Dietz (2020), the forced de-schooling and migration to a virtual education system that is monocultural seems to be a setback to the achieved progress of intercultural education from the last two decades. However, it may well be the opportunity to promote profound changes in that sense in the educational system. The same author states that something positive in this pandemic is the family role in education; grandparents, parents, aunts and uncles, brothers and sisters contribute to the educational achievement of children, adolescents, and young people, not only at home but also in productive work and in the community. However, it must be recognized that it has also led to an increase in the levels of domestic violence.

In this sense, Porlan (2020) goes to the extreme in his critical analysis of present educational systems given the current health crisis. He mentions that the opportunity lies in turning education with or without the pandemic, correcting the multiple deficiencies detected. He considers current education in most countries to be anachronistic, with urgency for substantial changes; it cannot continue to be a simple transmission of information from teacher to student. He suggests the need to exercise the mental activity of the subject and that the contents should be relevant, meaningful, systemic, and with a connection between the different disciplines.

The truth is that this health crisis can be an opportunity to promote educational proposals that seemed forgotten, challenging to accept, and even idealist. They can boost the integral formation of human beings more sensitive to nature, to the cruciality to make changes to mitigate the impact of the crises caused by humans, that had not been visible due to the

subjugation of the development models of an economist cut that privilege growth and capital accumulation.

Education, mainly higher education, must contribute to face the crises that we already had even without a pandemic: climate change, depletion of natural resources, hunger, and ungovernability in some regions. Given the uncertainty of the health crisis, there are at least three scenarios: an exaggeration of the crises, and therefore inequity and ungovernability, to take advantage of this situation; substantial changes generated, from which we all benefit; or an unfortunate return to "normality" as if nothing had happened.

As documented by Tzoc in Guatemala, Vallejo in Ecuador, Dietz in Mexico, and Hooker & Castillo in Nicaragua: the knowledge of indigenous communities, the ancestral wisdom inherited by generations, the respect for nature, including epidemics as a response from Mother Earth to an imbalance, and the resilience to face extreme challenges to which they have been exposed to, all offer us alternatives that should not be underestimated, but that we should listen to, through our students who come from these native communities.

Material and method

The qualitative approach study, based on the narrative and biographical-narrative research methodology, seeks to generate in the subject states of reflection and awareness of the experiences lived, as well as to take advantage of subjectivities to identify and explain the educational processes (Landín and Sánchez 2019). Stories and narratives give meaning to people's lives, and the experience of whoever intervenes in these processes becomes meaningful to the extent that there is a voice that is heard (Sparkes and Devís 2007). The systematisation of experiences can generate new knowledge, support appropriate individual and collective decision-making, and thus transform practice and context. In the studies on youth and conflict by Pinilla and Lugo (2011), young people express their experiences, share their world from their social and cultural constructions, from the complexity of their contexts. Just in these contexts, when they are subject to external pressure due to conflicts, the crisis can serve as an engine to generate initiatives, solutions to problems and creatively create support networks. It is the case of young university students at UIET.

This research reveals the students' ideas, experiences and concerns about the environment where they live; it shows how they face the health crisis, the risk, the forced isolation, and, in some cases, the grief, death, and illness of a family member due to the pandemic. For the collection of information, a research instrument was applied randomly to 70 students of the seven degrees offered by the UIET; 10 items were analysed, seeking to understand the perception about the changes in the process of their professional training during this period of distance work derived from the confinement of the COVID-19 pandemic. This work offered the student the opportunity to show their achievements, learning, challenges and experiences that are suddenly forgotten and, nevertheless, may represent relevant information for decision-making.

Results

Systematizing this profound and complex experience means putting the main actors of the educational process, that is, the students, at the centre of the analysis. The intention is to listen to their voice and to their feelings, how they lived this abrupt episode that changed the patterns of their university life. In the students' narratives, they have expressed their experiences in the educational process, isolated from their classmates, teachers and classroom. This pandemic has brought to light the shortages, the shortcomings of the conditions under which students from rural and indigenous regions live to complete their higher education, as shown in their answers to the questions posed. In one of them, they indicated their main lessons learned from this period:

> "I learned how to cope with problems; working online has been very good as it improves the quality of communication; there were problems, but I was able to cope with them." (Student 1. 2020)

As we can see, a positive aspect is that this pandemic led them to face the problem; it was an unexpected challenge. This comment indicates that the student liked this kind of exercise, working online. Even this discourse represents a great encouragement for the teacher, or as the following narrative that glimpses a certain autonomy in the study:

> "Not waiting for the teacher to solve or send the necessary information to carry out some of the activities. Profitably organizing the routine to allow time to complete the various activities." (Student 2. 2020)

It leads us to think that he had the resources to establish online or telephone communication because he saw it as something positive, despite the adverse situation; he did not stop in the face of the problem, but he saw it as an area of opportunity.

On the other hand, some expressed the opposite; the heaviness to carry out these school activities to what they were not used to, to see each other, to dialogue face to face:

> "A little complicated since the communication between teachers was very regular. However, there were times it was not very understandable. But despite the pandemic and distance education, we were able to finish our semester." (Student 3. 2020)

This note shows that from one side the student endeavoured, but the other part should come from the tutor, who had to find the best educational tool to be able to generate the student's interest and that, of course, it was the individual attention that, although it was exhausting, also allowed to know the student better, since he worked alone and the teacher could see his limitations, educational and personal technologies more clearly.

Another way of living this academic setback was what happened with those students who did not have access to technological resources and had less experience in the management of virtual resources; though the same happened with some of the teachers, the difference is that teachers had more possibilities to invest in a technological device.

> "The truth is that for me it was somewhat complicated and stressful, since I hardly knew how to handle some programs and I didn't have my computer, and the community where I live is very far from the village and doesn't have much of a signal, let alone a cyber." (Student 4. 2020)

It was not only the problem of whether they had a computer or not, but rather the problem of access to Internet service. Geographic characteristics highlight the limitations of external factors that prevent students from responding to virtual academic demands.

> "There were some complications with some of the topics to be covered, since in some cases the teacher had only given us what we had to research, without any additional explanation. Also, I could see that some classmates were not able to interact among us, sometimes, due to connection difficulties." (Student 5. 2020)

The diversity of perceptions is evident here; this judgment places the teacher's role in the spotlight. It refers to the teacher's discernment to think about the methodological process of how he guides the student in his learning. In-person, classes allow restating a question, suggestion, or opinion, but in this situation, how is it possible? In addition, the student appeals to interaction, a fundamental element in the educational processes that allow collective growth, but, unexpectedly, this form of schoolwork was restricted.

Another reason found in this exercise was the personal burden, in addition to the academic one:

> "The truth is, it was burdensome for me, since I am on my own and I didn't even have time to do my homework. I didn't hand over some assignments, I think it affected me. I just did what I could." (Student 6. 2020)

The activities that had to do with other responsibilities implied by the confinement were precisely those related to providing food, attending to home daily life, situations considered not decisive elements, and perhaps assessable factors of the learning process.

Despite everything, the students' responsibility was also nuanced in their comments since the goal was undoubtedly to obtain a positive grade to continue their academic training:

> "In general, this semester was a little complicated for me because of the lack of internet, and also because I live in a community that has poor phone signal and I had to spend money to recharge and keep up with my homework, but I still did my best to send my work timely and properly to do well in the semester." (Student 7. 2020)

Here we notice the voice of economic sensitivity in investing in a resource to fulfil with school activities. In this story, the student expressed the way she lived this episode:

> "Personally, it was a complicated task, given the different roles that I play until today as a member of a large family, where we have been affected by the virus, and logically this pandemic horribly dragged the household economy, leaving at the expense of many things. Later, the pressure of some activities in strict schedules was a factor to getting low grades. Besides, it is worth mentioning that there were times when I did not have internet and did not have access to an open cyber." (Student 8. 2020)

These data corroborate the conditions in which virtual classes were faced. Those who had the necessary resources to attend to the activities assigned by the teachers were the ones who were successful and passed the test. Those who did not have the required means, suffered the stress of not answering timely and efficiently as required by the adaptation of the virtual system to conclude the semester.

> "I had many difficulties concerning how I had to prepare my work [...] going outside home didn't help at all. But I looked for a way to do my work and, above all, to save my subjects. In fact, the most convenient thing is to be in a classroom where they can explain in detail." (Student 9. 2020)

On one side, the whole mechanism that moves the student to face the virtual modality in this pandemic becomes evident. From those who can do it, because they have the economic, material, and the family support, to those who lack all the means to succeed in these tasks. What is the point of lockdown if they have to leave home to go to a cyber and stay the necessary time to upload their homework or do their research? The teacher does not realise this; he is interested in how the student answers to turn his progress. On the other hand, this voice longs for the in-person way of studying; it reflects the need for a teacher to support their learning.

> "[...] this semester that already finished was [...] very stressful because you don`t gather with other people to share ideas and get experience with them as participations and dialogues, it was a little complicated because the internet equipment was missing, although it was very forced to make the hiring, but anyway I hired it for me to do my activities so that I would not fail the semester [...]." (Student 10. 2020)

It evidenced the feeling of responsibility for the academic load, which turned into stress. In addition, it incurred in an expense that was not within the plan, to hire internet. All this is a minimized factor, while some students appeared with failing grades. As we can see, several of these factors were not considered as decisive elements in the fulfilment of the tasks and the evaluation of the academic performance of the students at UIET.

Conclusions

The research results show the great responsibility that higher education has for the transformation of Latin American societies. Especially, intercultural university education provided in regions with ethnic and cultural diversity must build spaces for coexistence in diversity, with respect for identity, human rights, and local knowledge. It can be concluded that the presence

of COVID-19 and prolonged confinement have had significant effects on the teaching-learning process of university students, especially in rural and indigenous areas, which had their academic performance negatively affected and, in some cases, their possibilities of continuing and completing their studies.

It is essential to consider the need to modify learning strategies, content development and evaluation indicators. That includes not only the knowledge, skills and attitudes specific to their disciplines, but also those more general that allow them to coexist in cultural diversity, especially in times of crisis.

There are crucial limitations for the students in rural areas for their learning process: the inequality reflected in poverty, marginalization and lack of means for the study. Hence will affect the results of academic evaluation. In addition, it is predicted that there is a risk of psychological sequels due to prolonged confinement and high stress and frustration due to the limitations to comply with the evaluable products. Therefore, it should be considered in the tutorial and psycho-pedagogical attention. However, valuable lessons came from this experience, which will be shared with the student community, teaching staff, and directors, to continue facing the problem of academic activities in lockdown, since it has not yet concluded and threatens to continue in the next school year.

A positive aspect to consider is that reports of experiences, new knowledge, strategies, and, in general, new ways of facing the problem and working on this pandemic have generated rapidly, worldwide, something that can be used for analysis and decision-making by the teaching staff. Now, more than ever, the importance of face-to-face activities and classroom participation, typical of the intercultural educational model, have been valued. So, strong doses of creativity will be required so that this environment can continue through the electronic media.

Finally, we agree with several of the authors quoted here. It is urgent to modify educational institutions and policies not to return to “normality”, but to open new educational paradigms which will substantially improve access to higher education and include intercultural education as a desirable quality. The challenge is to fully comply with the human right to quality and inclusive education.

Bibliography

Acosta A. 2020. "La educación superior ante el COVID-19. Un nuevo reto y viejos resabios". *Reporte CESOP. 2020. COVID-19, la humanidad a prueba. Número 132, edición especial, mayo 2020.* México. Cámara de Diputados. http://www5.diputados.gob.mx/index.php/camara/Centros-de-Estudio/CESOP/Tema-Covid-19/Reporte-CESOP.-Covid-19-La-Humanidad-a-Prueba.-Edicion-Especial.-Num.-132-mayo-2020.

ANUIES. 2020 a. "Hacia la construcción colectiva de la nueva normalidad en la educación superior". México. ANUIES. https://educacionsuperiordurantecovid.anuies.mx/wp-content/uploads/2021/04/Planeacion_de_un_regreso_seguro.pdf.

ANUIES. 2020 b. "Sitio de Espacio Docente": https://espaciodocente.mx/index.html.

Balluerka L.N., Gómez B.J., Hidalgo M.M.D., Goroztiaga M.A., Espada S.J.P., Padilla G.J.L., y Santed G.M.A. 2020. "Consecuencias psicológicas del COVID-19 y el confinamiento". España: Universidad del País Vasco.

Buendía E. A. 2020. "Desafío de la educación superior en tiempos de pandemia: la contingencia inesperada". *Reporte CESOP. 2020. COVID-19, la humanidad a prueba. Número 132, edición especial, mayo 2020.* México: Cámara de Diputados. http://www5.diputados.gob.mx/index.php/camara/Centros-de-Estudio/CESOP/Tema-Covid-19/Reporte-CESOP.-Covid-19-La-Humanidad-a-Prueba.-Edicion-Especial.-Num.-132-mayo-2020.

Castillo A. M. 2020. "Opciones para la actividad escolar durante la pandemia en COVID-19 el caso de Nicaragua, aportes para enfrentar la pandemia". *Serie Ciencia, Técnica y Sociedad*. Nicaragua: Academia de Ciencias de Nicaragua.

Cendales G. L. y Torres C. A. 2006. "La Sistematización como experiencia investigativa y formativa". *Revista La Piragua. No. 23.* http://www.cepalforja.org/sistem/documentos/lola_cendales-alfonso_torres-la_sistematizacion_como_experiencia_investigativa_y_formativa.pdf.

CEPAL. 2020. "América Latina y el Caribe ante la pandemia de COVID-19. Efectos económicos y sociales". *Comisión Económica para América Latina y el Caribe (CEPAL). Informe Especial, numero 1.*

CONEVAL. 2018. "Medición de la pobreza 2008-2018". *Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL)* México. https://www.coneval.org.mx/Medicion/MP/Paginas/Pobreza-2018.aspx.

CONEVAL. 2020. “La política social en el contexto de la pandemia por el virus SARS-CoV2 (COVID-19) en México”. *Consejo Nacional de Evaluación de la Política de Desarrollo Social (CONEVAL)*. México.

Cotino H. L. 2020. “La enseñanza digital en serio y el derecho a la educación en tiempos de coronavirus”. *Revista de Educación y Derecho, número 21, octubre 2019-marzo 2020*. España: Universidad de Valencia.

Dietz G. y Mateos C. L. S. 2020. “La interculturalidad educativa en tiempos de pandemia. Muchas sombras y algunas luces”. *Educar en la Diversidad. Boletín del Grupo de Trabajo Educación e Interculturalidad. CLACSO*.

EDUL@Ab. 2020. “Decálogo para una docencia online inesperada”. *Universidad Oberta de Cataluña.* http://edulab.uoc.edu/es/2020/03/16/decalogo-para-docencia-linea-inesperada-2/

Failache E., Katzkowicz N. y Machado A. 2020. “La educación en tiempos de pandemia y el día después: El caso de Uruguay”. *Revista Internacional de Educación para la Justicia Social. 2020, 9(3e) número extraordinario “Consecuencias del cierre de escuelas por el COVID-19 en las desigualdades educativas.*

FAO. 2020. “Seguridad alimentaria bajo la pandemia de COVID-19”. *Organización de las Naciones Unidas para la Alimentación y la Agricultura y Comunidad de Estados Latinoamericanos y caribeños.* http://www.fao.org/3/ca8873es/CA8873ES.pdf.

Gazzo M. F. 2020. “La educación en tiempos de pandemia COVID-19, nuevas prácticas docentes ¿nuevos estudiantes?” *RED Sociales; Revista de Departamento de Ciencias Sociales. Volumen 07, numero 02.* https://ri.unlu.edu.ar/xmlui/bitstream/handle/rediunlu/750/9.-La-educaci%C3%B3n-en-tiempos-del-COVID-19_-nuevas-pr%C3%A1cticas-docentes-%C2%BFnuevos-estudiantes_.pdf?sequence=1&isAllowed=y.

Hooker B. A. y Castillo G. L. 2020. “La educación superior en contextos multiculturales, visión de futuro”. *Revista Ciencia e Interculturalidad. Año 13 volumen 26 numero 1.*

Landín M.M.R. y Sánchez T.S.I. 2019. “El método biográfico – narrativo. Una herramienta para la investigación educativa”. *Revista Educación. Volumen XXVIII, Número 54. Universidad Veracruzana. Pgs. 227-242.*

Mérida M. Y. y Acuña G. L.A. 2020. “Covid-19, Pobreza y educación en Chiapas: Análisis a los programas educativos emergentes”. *Revista Internacional de Educación para la Justicia Social. 2020, 9(3e).*

OIT. 2020. "Los jóvenes y la COVID-19: efectos en los empleos, la educación, los derechos y el bienestar mental". *Organización Internacional de Trabajo (OIT). Informe de la Encuesta 2020.*

R. 2020. "El cambio de la enseñanza y el aprendizaje en tiempos de pandemia". *Revista de Educación Ambiental y Sostenibilidad 2(1) 1502*. España. Universidad de Cádiz.

Ramírez-Ramírez L. N., Arvizu-Reynaga V., Ibáñez-Reyes L., Claudio-Martínez C. y Ramírez-Arias V. 2020. Apoyo ante COVID-19 en Latinoamérica, estudio exploratorio de las necesidades, psico-socio-educativas durante la contingencia". *Congreso Internacional Virtual sobre COVID-19, consecuencias psicológicas, sociales, políticas y económicas.*

Reimers F. M. y Schleycher A. 2020. "Un marco para guiar una respuesta educativa a la pandemia del 2020 de COVID-19". *Organización de Estados Iberoamericanos* (OEA). https://www.oei.es/Ciencia/Noticia/la-oei-difunde-un-informe-de-la-universidad-de-harvard

Pinilla S.V.E. y Lugo A.N.V. 2011. "Juventud, narrativa y conflicto. Una aproximación al estado del arte de su relación". *Revista Latinoamericana de Ciencias sociales, niñez y juventud. Volumen 9, numero 2. Julio-diciembre 2011*.

Renna G. H. 2020. "El derecho a la educación en tiempos de crisis: alternativas para la continuidad educativa". *Universidad Abierta de Recoleta, Universidad Nacional Experimental Samuel Robinson. Clúster de Educación*. Hooker B. A. y Castillo G. L. 2020. "La educación superior en contextos multiculturales, visión de futuro". *Revista Ciencia e Interculturalidad. Año 13 volumen 26 numero 1.* Venezuela, Caracas.

Sparkes A.C. y Devís D.J. 2007. "Investigación narrativa y su forma de análisis: una visión desde la educación física y el deporte". http://viref.udea.edu.co/contenido/publicaciones/memorias_expo/cuerpo_ciudad/investigacion_narrativa.pdf.

Tzoc J. 2020. "Multidimensionalidad en el pensamiento de los pueblos en torno al coronavirus". *Universidad Rafael Landívar de Guatemala. Revista Científica Internacional.*

UAM. 2020. "El Proyecto de Enseñanza Remota de la UAM, una medida temporal por el COVID-19". *Boletín UAM, número 271, 11 de Mayo de 2020.* https://www.uam.mx/ss/s2/comunicacionsocial/boletinesuam/271-20.html.

Vallejo I. R. y Álvarez K. 2020. "La pandemia del coronavirus en la Amazonia Ecuatoriana. Vulnerabilidades y olvido del estado". *Cuadernos de Campo. Sao Paulo online. Volumen 9 numero 1.*

Vijil J. 2020. "La educación en Nicaragua: Emergencia más allá del COVID-19. En *COVID-19 el caso de Nicaragua, aportes para enfrentar la pandemia*. Serie ciencia, técnica y sociedad. Nicaragua. Academia de Ciencias de Nicaragua.

Villamizar E. J. D. y Barbosa C. J. W. 2017. "Sistematización de experiencias. Indicadores y elementos representativos para la investigación educativa". *Revista Espacios. Volumen 38 número 47. Año 2017*.
http://www.revistaespacios.com/a17v38n47/a17v38n47p26.pdf.

Vivanco Á. A. 2020. *Teleeducación en tiempos de COVID-19: brechas de desigualdad.* CienciAmérica volumen 9 (2).

CHAPTER VIII

DISCOVERING RELATIONS FOR ESTIMATING LENGTH OF THE HOSPITALIZATION TIME OF COVID-19 PATIENTS: A FIRST APPROACH

GEMAYQZEL BOUZA-ALLENDE[1],
ELA M. CÉSPEDES MIRANDA[2],
ROLANDO J. GARRIDO GARCÍA[2],
ROGER RODRÍGUEZ-GUZMÁN[2],
PABLO SOSA PEDRO[3] AND
NIURELKIS SUÁREZ CASTILLO[2].

Summary

During the pandemic outbreak, the number of patients in hospitals increased dramatically in a short-term period. This stressed the capacities of the hospitals. The goal of this work is to investigate which variables can be used to estimate the number of days a patient can be at a hospital. Using data of patients recovered from COVID-19, correlations between the number of days at a hospital, personal characteristics of the patients and previous diseases, symptoms and treatments have been studied. A classification of patients using K-means was obtained.

Keywords: Correlations, COVID-19 recovered patients, K-means.

[1] Universidad De La Habana, Cuba.
[2] Universidad De Ciencias Médicas De La Habana, Cuba.
[3] Policlínico Reinaldo Ti Mirabal, Cuba.

1. Introduction

The outbreaks of COVID-19 have had a lot of consequences worldwide. Several health systems have collapsed, due to the high number of patients that needed medical treatments and resources at the same time. So, it is important to estimate how long a certain patient may stay at the hospital and the services he/she will need.

Since the first outbreak, it was clear that patients with certain diseases would probably have a more complicated evolution. For instance, in a study of Wuhan patients, see [32], the authors found that the potential risk factors of a poor diagnosis in elderly, diabetic and high arterial tension patients were age, high SOFA score and D-dimer values. A similar study was carried out with the members of the UK Biobank Community Cohort, see Atkins et al. [3]. This study showed that dementia, diabetes, chronic pulmonary disease, pneumonia and depression were statistically significant related to mortality. Moreover, in Ssentongo et al. [26], the statistical study on 65,484 patients from China, the United States, Italy and South Africa hints that individuals with cardiovascular disease, hypertension, diabetes, congestive heart failure, chronic kidney disease and cancer had more risk of dying. According to these statements, [24] emphasizes that, although the probability of getting infected with Coronavirus is the same for cancer patients and healthy people, lethality in the cancer patient group is higher and statistically significant (32.1 % vs 3.5%). [5] applied a cluster technique to predict the diagnosis of elderly infected with COVID-19, based on their characteristics and diseases that affected them. The global study, presented in [28], suggests that patients with renal, cardiovascular or cerebral-vascular diseases had more COVID-19 severity and a larger risk of mortality.

Other characteristics of the evolution of the patient would be interesting. From the viewpoint of the medical logistics, see [18], a forecast of the length of hospital-staying is desired. A first attempt based on an estimation of the probabilities of staying at the hospital can be found in [2]. However, a more accurate estimation based on other factors is required. Some questions that can be derived are:

- May characteristics such as sex, age and previous diseases provide appropriate estimation of the expected time the patient may stay at the hospital?
- Is the estimation more accurate if the symptoms are also included?

- If there are different treatments, is the recovery period different for similar groups with different treatments?

Analogous questions can be formulated to estimate what kind of diseases may appear after the end of the infection. A study related with the appearance of sequels suffering COVID-19 can be found in [15].

Some studies have proposed mathematical models to support medical decisions based on prediction of length hospital staying. Motivated by the outbreak of COVID-19 and the need of efficiency in other health services, [1] studied through a Cox prediction model the relation between the days at the hospital by diseases such as leukaemia, kidney and heart diseases. In [20], a model based on decision trees has been proposed to estimate the length of the hospital stay of COVID-19 patients after analysing 2,017 individuals in Dubai. [30] proposed estimations of this variable based on three approaches: an accelerated failure time survival model, a truncation corrected method and a multi-state survival model. The study was done on a database of 6,208 English COVID-19 patients.

In this contribution we illustrate the use of classification techniques for estimating the length of the hospital-staying (LHS) in surviving individuals. To perform the estimation, we used a sample of Cuban recovered patients living in Havana. The interviews were performed in March-June 2021 with individuals that got infected between April 2020 and May 2021. For each patient, we collected socio-demographic characteristics, the diseases before the infection, the treatments the individual was following, the reported symptoms, the number of days at the hospital and the required treatment. We want to point out that during the studied period, in Cuba, a patient diagnosed by a PCR test was immediately admitted at the hospital. Release could take place just after PCR turned negative.

We divided the sample into two parts: the training sample and the test sample. With the data of the training sample, we considered different combinations of indicators. The LHS was always included, and a K-means cluster classification was performed. For more details on the algorithm and their implementation, see [13, 31, 27]. The individuals in the test sample were classified into the obtained clusters. With this aim, all the indicators, except the LHS, were considered. The difference between the true value of LHS and the estimation provided by the centroid of the cluster was used for evaluating the classification error. The study was complemented with a

correlation study between the indicators, and the comparisons of the mean value of the LHS among different groups of interest.

The contribution is organized as follows: First, we present how the data were organized. In Section 3 and 4 we present the correlations obtained between the collected variables and the tests, for determining the existence of a significant difference between the means of LHS for different values of the other variables. Section 5 includes the cluster analysis. The article ends with some remarks and the proposal of future lines of research.

2. The data

In this section we present the collected data and explain the performed data analysis. Doctors visited a selected set of patients from three health sector belonging to three municipalities of Havana city. After expressing his/her willingness to participate in the study, an interview was conducted for collecting the needed data.

The collected data were organized considering the dynamics of the disease. As the sample included only recovered patients, everyone followed the sequence of susceptible-infected-recovered states.

The first group corresponds to the Personal Data of the Patient (PDP). In this set we included Age (in years), Sex (females=0. male=1), Colour of the Skin (white=W, mixed=M, black=b) and Blood Type (O-,O+,A+, A-, B+, B-, AB+, AB-, unknown). Group H has the information about the health status of the patient before getting infected. We considered three sub-groups: the diseases they suffered, the associated medical treatment and other therapies the individual followed. The categories of diseases were Hypertension (HT), Allergies, Asthma, Diabetes Mellitus (DM), Ischemic Heart Disease (IHD), Cancer, Dengue, Smoker, Other Diseases. For each disease, and each individual I

$$H(I,j) = \begin{cases} 1, & \text{the individual } I \text{ has the disease j} \\ 0, & \text{otherwise.} \end{cases}$$

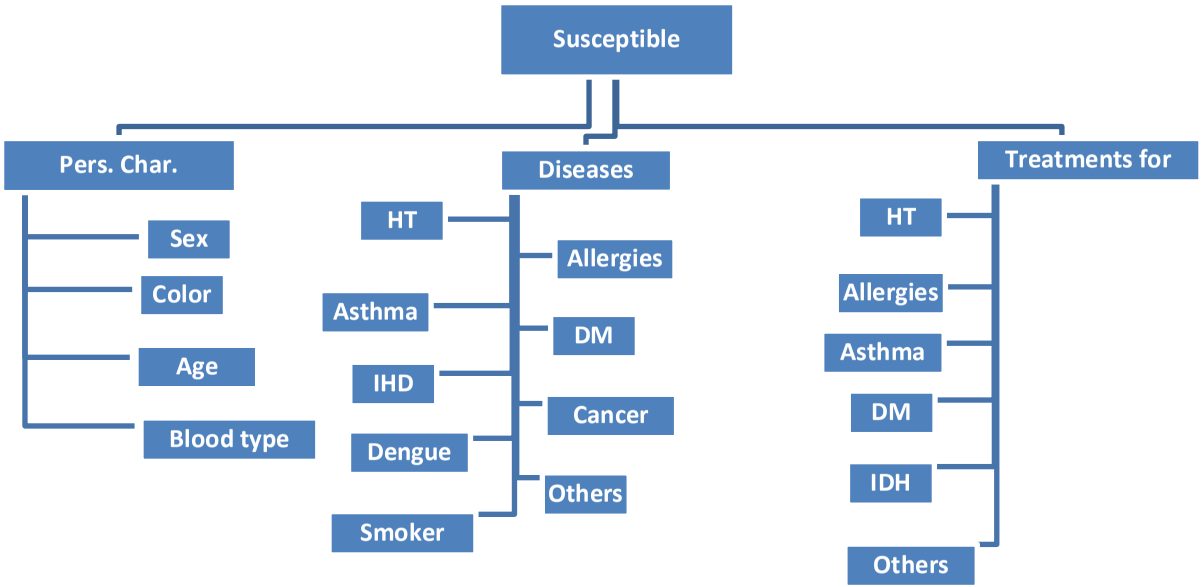

Susceptible
Pers. Char.
Sex
Color
Age
Blood type
Diseases
HT
Allergies
Asthma
DM
IHD
Cancer
Dengue
Others
Smoker
Treatments for
HT
Allergies
Asthma
DM
IDH
Others

Figure 1

For the treatments, we considered those corresponding to the diseases mentioned above and others. The 1-0 convention used for the diseases was considered. PDP and H are the groups corresponding to the characteristics of the patient during the susceptible period. Roughly speaking, these data are organized as depicted in Figure 1.

The data related with the infection period is divided into three groups: the symptoms, LHS and the prescribed therapy. Of course, in the first group many combinations may appear. So, the most common symptoms such as headache, fever, fatigue, loss of taste and smell are considered as separate categories. The category "others" corresponds to rare symptoms like high tension and pain behind the eyes. We want to point out that in the variable Fatigue we included the symptoms decay and weakness. The Cuban protocol in that period included supplying Herferon, Interferon and Heberferon to the infected individuals. These medicaments are boosters of the immunological system. Each patient received one of them in several doses. This number was not collected because most of the patients did not remember how many times they got the medication. In general, it depended on the length of the infection. For possible complications, the most used treatments are grouped in the variables Antibiotics and Chloroquine and Kaletra (CK). Figure 2 shows how the data corresponding to symptoms and treatments were organized.

The idea is to estimate the LHS using information of the group's PDP, H and D. Here, D represents the data related to the symptoms and the treatments of the disease.

We have used three main statistical techniques: correlation, mean comparison and cluster analysis. Pearson and Spearman correlation indices have been considered. Using a t-test, we analysed if the mean of the LHS was different for groups such as males and females, patient with a certain disease (being under a certain treatment) or not. For the clusters, we have considered the combination of data of the form [LHS, C], where C is a set of variables belonging to PDP, H and D.

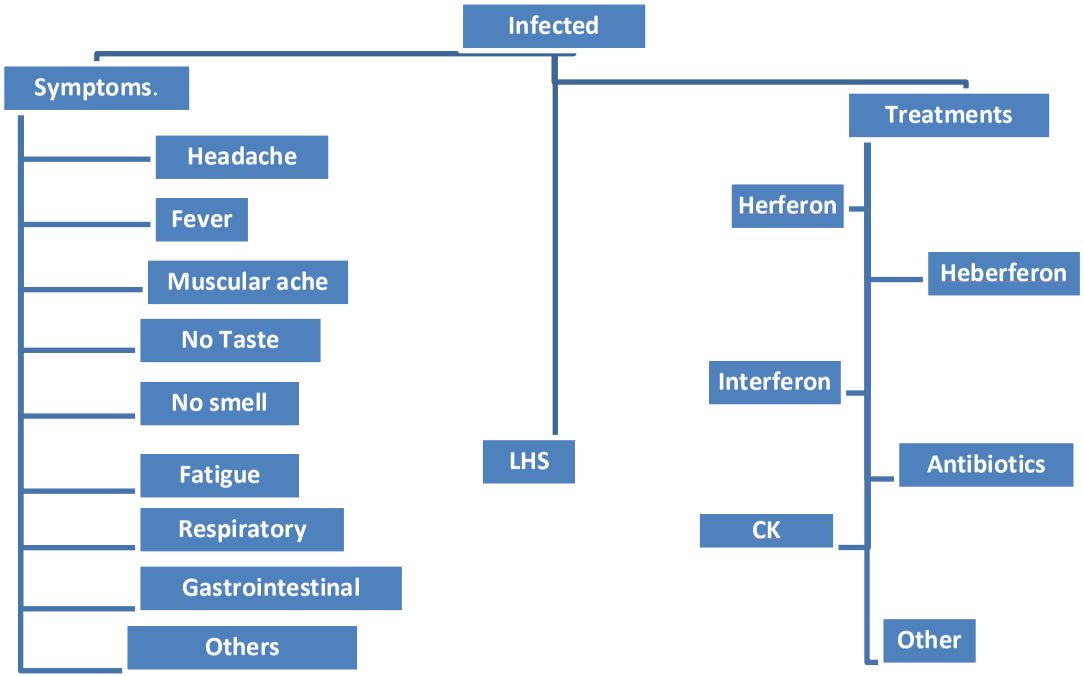

Infected
Symptoms.
Headache
Fever
Muscular ache
No Taste
No smell
Fatigue
Respiratory
Gastrointestinal
Others
LHS
Treatments
Herferon
Heberferon
Interferon
Antibiotics
CK
Other

Figure 2

The LHS has always been included, as it was used for testing the classification. The sample was split into two parts: a training and a test sample. Fixed C for the combination [LHS, C], the clusters were computed using the training sample and K-means as clustering technique. Then, each individual I=1, …, N, of the test sample was assigned to the cluster whose centroid was closer to I according to the values of C. If I is assigned to cluster J(I), the evaluation of the classification is given as E(I), the error between the LHS of the centroid of cluster J(I) and I. The best combination will be such that [E (1), ..., E(N)] is smaller.

All the statistical analysis was performed using IBM-SPSS-20. The level of significance used in the statistical tests is 0.05.

3. Statistical analysis of the relation among PDP, H variables and Length of the Hospital Staying (LHS)

We will illustrate the proposed methodology with a sample of 282 post-COVID-19 individuals living in Havana. For the training set, we considered 229 patients, being 121 women and 108 men. From the total patients.128 had white skin, 59 mixed and 42 black. The age mean was 44.4 years old with a standard deviation of 20.1 and quartiles 27, 45 and 57. The test set had 53 elements, 31 were women and 22 men; 39 had white skin, 4 mixed and 10 black. On average, they were 40.2 years old with standard deviation 20.2. The quartiles are 24 42 56. So, the distribution of the age in the two groups is similar. The proportion between women and men in the test set (1.41) was larger than in the training set (1.12). Similarly, the distribution of people by colour of the skin was quite different: the percent for white, mixed and black colour of skin was 56%, 25.7% and 18.3% for the training set, and 73.6%, 7.5% and 18.9% for the test set. However, this difference seems not to be important. More details can be found in Figure 3, where the mean, and quartiles of the age are displayed for each group.

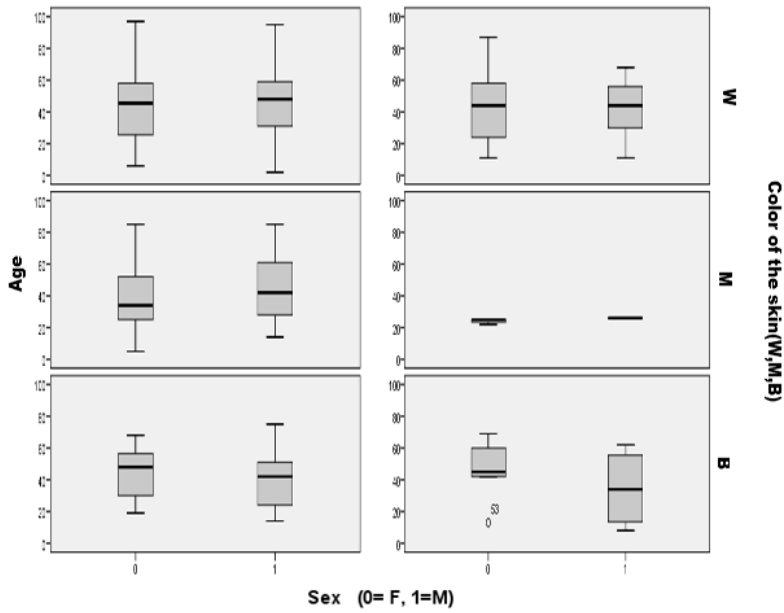


Figure 3

The goal of this paper is to estimate the number of days the patient shall stay at the hospital. We have obtained that LHS shows no significant differences for people of different sex and/or different colours of the skin.

3.1. Correlations

We have studied the correlations among the variables in the group's PDP and H. We have analysed, for each variable V in PDP and H, if different values of V imply the existence of a significant difference in the mean of the LHS. As Age and LHS were numerical variables and the others were dichotomous, we have used the Pearson coefficient

$$r = \frac{N\sum_{n=1}^{N} x_i y_i - \sum_{n=1}^{N} x_i \sum_{n=1}^{N} y_i}{\left(N\sum_{n=1}^{N} x_i^2 - (\sum_{n=1}^{N} x_i)^2\right)(N\sum_{n=1}^{N} y_i^2 - (\sum_{n=1}^{N} y)^2)}$$

for computing the correlation between them, see [11,12].

In the sample, a significant (negative) correlation was observed between Sex and the variable Other Diseases. Similarly, we obtained that Colour of the Skin was positively correlated with allergies and IHD, the darker coloured individuals showed more propensity to allergies and IHD. However, it shall be pointed out that a large percentage of the white skinned individuals reported suffering HT.

We also obtained that the variables Age, DM and HT were significantly positive correlated among them. This is expected since in Cuba, in [22], it is stated that DM, HT and its complications (IHD) are more common in elderly. The positive correlation of Age, Cancer, IHD and Other Diseases is also significant. In [4], it is reported that Cancer and Aging share many hallmarks such as genomic instability and the same strategies, and drugs can be used to target both, so our association is justified. The relationship between Age and the appearance of IHD and Other diseases is also well known.

A negative correlation between Age and Allergies was observed (allergies were mostly reported by younger people). This may be a consequence of the observation found in [7]. The authors reported that most of the allergy's studies are done in children rather than in elderly, so misinformation/ misreport could be an important element for allergy in older population. Another consequence of the fact remarked in [7] could be the observed negative significant correlation between suffering HT and Allergies. In particular, the incidence of HT is higher in aged individuals. Furthermore, in Cuba, HT together with other cardiovascular diseases are reported as the major cause of mortality in elderly, see [22]. On the other hand, a higher effectiveness of treatments for decreasing Systolic Blood Pressure has been observed in patients with appropriated allergies treatments, see [19].

Positive association between being a Smoker and IHD was also obtained. This positive association can be justified as follows. The Framingham Heart Study, the most important epidemiological study for cardiovascular diseases, see Burke et al [6], have observed the existence of a relationship between smoking and IHD. New evidence reinforces this fact. In [25] it is reported that smoke causes endothelial dysfunction, and this dysfunction is related with the appearance of IHD, see [23].

We also observed the collateral positive correlation between suffering from Cancer and having been affected by Dengue. [9] reported that patients with dengue had a higher risk of leukaemia. As far as we know, association

between dengue and other types of cancer have not been studied or are not available.

In almost all cases, a variable significantly correlated with a certain disease is also correlated with the corresponding treatment. This is an expected result, due to the full coverage of the Cuban health system, that allows to follow up and treat these diseases. The only exception is in the case of the correlation of Other Diseases and Other Therapies with Sex. A reason for this result is that many healthy people report they follow other therapies such as vitamins and immunological boosters.

With respect to the blood type, we have only considered individuals of groups O+, A+ and B+ (the number of observations in the other groups was too small and 82 patients did not know their blood type). The variables of the PDP and the H groups showed no significant difference across the considered three groups. For instance, we cannot reject the hypothesis the proportion of HT in groups O+ and A+ is equal (p=0.6).

Then, we considered the correlation of these variables with LHS. As expected, a positive correlation between this factor with Age, Diabetes and IHD was observed. The rest of the variables of the groups PDP and H did not show significant correlation. These risk factors were also observed in [8]. In [29], correlations between LHS and the characteristics of the individual were also carried out. They also observed a significant correlation between LHS and HT, which in our case was not significant (p=.158). However, this p-value was very small, compared with the significant level of the correlation between LHS and Asthma, Allergies, Smoking, Cancer and Other Diseases (p= .707, .407, .932, .802, .795).

3.2. Comparison of the mean of LHS for different groups of individuals

In this part we will show the study that analyses if there is a significant difference between the mean of LHS across the different groups defined by the rest of the nominal variables in PDP and H. Although the sample was not random, we have used the statistical tests of equality of the means as a clue of what it seemed to be going on. That is, we compared the mean of LHS across individuals of different sex, colour of the skin and different blood type. For the H data, we developed a similar analysis. Having fixed a disease of the categories given in H, we studied the mean of LHS for patients

having it or not. For Age, we found age values where the difference of the means of LHS were statistically different.

We obtained that significance difference of LHS between males and females is not to be accepted. The same holds for the variable Colour of the Skin. With respect to Blood Type, a significant difference of the means of LHS only appeared between the groups O+ and B+, individuals of the first group shall stay approximately 4 days more. As already mentioned, the analysis was performed only through the groups O+, A+ and B+.

The use of the t-tests sustained that the mean of the LHS was different for IHD patients. Indeed, on average, these patients stayed 6.372 days more in the hospitals. The variances cannot be assumed equal and with p=.022, so we can accept that the means are statistically different.

With respect to the rest of the diseases, we could not reject the hypothesis that the means of LHS are equal with p=0.05. Indeed, the larger difference (3.6 days) appeared between Diabetes patients and the rest of the sample, although as p=0.087 was not statistically significant. Larger p-values (0.158, 0.707, 0.407, 0.802, 0.426,0.795) were observed for people suffering HT, Asthma, Allergies, Cancer, Dengue and Other Diseases. No difference in LHS was also observed between smoking or not (p= .932). In [10], the relation between LHS and comorbidities was also studied. HT and DM patients have a significant larger stay at the hospital. However, in their sample, hospitalized individuals were considered and, as in most countries healthy people normally stay home, it was expected that HT and DM should have a larger incidence than in ours.

With respect to Age, we have performed an experiment to detect at which age the means of LHS for the group of older and of younger patients should be significantly different. First, we considered 20, 40, 60 and 80 years old as cutting points. The only case where the significance of the difference could be accepted was for 40 years (p=0.003). Analysing the interval (20,50), the ages 25, 30 and 35, 45 were considered. The best result, p=.001, was obtained for Age=30. Patients older than 30 should stay 2.6 days longer on average, and this difference was statistically significant. Compared with [17], on average LHS in Cuba is smaller. Dividing the sample considered in [17] as younger than 70 years old and older, the difference of the mean of LHS is 32 days. In Cuba this cutting point appears earlier, and the difference is smaller. This may be again a consequence of the protocol used in the studied period.

4. Relations between PDP, H and D variables and Length of the Hospital Staying (LHS)

In this part we are going to include the study of the relationship between LHS and D, specifically symptoms and treatments. We again considered the training sample, whose characteristics were described in Section 3.

4.1. Correlations

First, we analysed the correlation of the variables in D with those in PDP and D. As in Section 3.1, the Pearson correlation coefficient was considered. As a result, we obtained a negative correlation among Headaches, Loss of Smell and CK with Age. Older people presented more fatigue and Other Symptoms, as Age and Fatigue were significantly positive correlated. Certain symptoms and treatments were significantly correlated with the sex. Fever was more common for male individuals and loss of smell in women. More men used Herferon. Indeed, the probability of using Herferon for men is 0.1 and 0.01 for women, and, clearly, we could reject the hypothesis that they were equal ($p=0.003$). With respect to previous diseases, DM and HT are positively correlated with fatigue. Losing the smell is negatively correlated with having HT.

Different blood type individuals showed similar symptoms and no significant difference with respect to the used therapies. As a matter of fact, when testing the proportions of people of type O+, A+, B+ having a certain symptom, we cannot reject the hypothesis that these proportions are equal (in all cases, $p>0.09$). So, the appearance of symptoms does not differ if blood type is O+, A+ and B+.

Muscular pain is also positively correlated with headache, fever and fatigue, while Headache is positively correlated with all the other symptoms considered in group D. We also obtained a positive significant correlation among Loosing Smell and Fatigue and Loosing Taste. Fatigue is also correlated with fever and use of antibiotics.

Between the different medicaments used for treating the SARCOV-2 infection, we observed a positive correlation between the Antibiotics and the CK. Since most patients got one and only one of the medicaments Herferon, Interferon or Heberferon, the expected negative significant correlations were obtained.

Physicians expect longer hospitalization time if there are symptoms. As a result of this study, we obtain that this assumption was valid in the case of Fever and Fatigue, due to the significance of the positive correlation of these two variables with LHS. In our case, we observed LHS was correlated with more symptoms than in [14]. They only reported a significant correlation between LHS and fever. Our results differed from those reported in [29], where only respiratory symptoms were correlated with LHS.

In the case of treatments, we observed that only Antibiotics and CK were positively correlated with this variable. It is also an expected result because these therapies are prescribed when there are complications, and this fact is usually related with longer hospitalization time. No significant correlation of LHS and the three therapies used to booster the immunological system was observed. Recall that Herferon, Interferon or Heberferon are supplied to almost all the patients when the infection is detected. Indeed, only 14 patients did not have any of these medicaments. So, there are not enough data to determine if LHS is significantly larger if these treatments are not consumed.

4.2. Comparison of the mean of LHS for different groups of individuals

In this part for all symptoms, we will demonstrate a test performed for analysing if there is a significant difference of the mean of the LHS for individuals having symptoms and those without them. An analogous study was developed for each therapy.

For muscular pain, headache, respiratory symptoms, loss of taste and loss of smell, we cannot reject the null hypothesis ($p>0.1$, in all cases). For patients with fever, the mean of LHS was almost 4.5 days longer. In this case the null hypothesis is rejected with ($p<10^{-3}$). The Levine test, see [11], also leads to different values of the variance ($p<10^{-3}$). In the case of fatigue, the difference between having this symptom or not was 2.9. As in the previous case, we can reject the equality of the means and the variances of LHS of the two groups with $p=0.005$ and $p=0.002$, respectively. Similarly, we can reject the hypothesis that LHS is the same if there are gastrointestinal symptoms ($p=0.001$). On average, individuals with this disorder shall stay 3.73 days more at the hospital.

For the treatments, the equality of the means was not rejected for Herferon, Interferon and Heberferon (the differences were equal to 1.86, 0.311, 1.017,

respectively with corresponding p-values equal to 0.286,0.694, 0.172). However, the data was collected after a personal interview and many patients were not sure which medicaments they had received. With a more accurate information, a different result may be obtained. Smaller LHS's are expected when Heberferon is used. We want to point out that the variance of LHS for patients using Heberferon is different according to the corresponding Levene test (p=.026). In the first group the variance is 21.094 and in the second, 40.05.

The combination of Chloroquine and Kaletra implied that, on average, the patients that required this treatment will stay 3.6 days more at the hospital. As $p<10^{-3}$, the equality was rejected, and we can assume that the means are different. In the case of the antibiotics, the difference was 3.2 and p=0.019. Similarly, the equality was rejected.

5. Cluster analysis

As already mentioned, the goal of this work was to obtain an approximation of LHS. With this objective, we have considered combinations of variables including LHS and compute the clusters applying the K-means procedure to the training set, see [13]. Given the centroids, we classified the test groups: for each element we have considered the vector of variables without LHS. Then, we computed the distance of the reduced vector to each (also reduced) centroid. Each element was associated to the cluster defined to the nearest centroid. An element would be well classified if the difference of LHS associated to the centroid (LHS_{app}) and the actual LHS (LHS) of the individual was sufficiently small. We have considered the following combinations:

- [LHS, S, T]: length of hospital staying, symptoms and treatments
- [LHS, S, T, Age]: length of hospital staying, symptoms, treatments and Age
- [LHS, S, T, HT]: length of hospital staying, symptoms, treatments and HT
- [LHS, S, T, Age, HT]: length of hospital staying, symptoms, treatments, Age and HT

- [LHS, S, T, Age, H]: length of hospital staying, symptoms, treatments, age and H data

- [LHS, S, T, PDP, H]: length of hospital staying, symptoms, treatments, PDP and H data

- [LHS, S, T, PDP, HT]: length of hospital staying, symptoms, treatments PDP and HT

Other combinations were considered as uninteresting. As a matter of fact, these variables had the same value across the clusters. Indeed, in the case of H variables, only HT introduces variability. When the other diseases are included in the construction of the clusters, all the centroids include healthy individuals. With respect to the other PDP data, there are differences, but as we will see there is not a large difference in the clusters, defined by all the data of PDP, and those obtained when only the age was considered. We applied the K-means procedure to compute clusters using the variables in the group's PDP and H, D and LHS. The use of the option provided by Hierarchical Cluster proposed to compute 4 clusters.

5.1. Computation of the clusters using the training sample

We have used the training sample of 229 individuals described in Section 3. The clusters computed by the combination [LHS, S, T] are represented in Table 1.

If HT is added, the same centroids are obtained.

As we can observe, at the centroids of the generated clusters, the values of the variables Herferon, Interferon and Heberferon are the same. As we have already mentioned, many patients did not provide an accurate information on which kind of immunological booster they had received. Shorter LHS are related with no symptoms. The centroid of the cluster with 25 days of LHS also includes having symptoms of fever and fatigue and Other Treatments. So, most of the elements of the cluster had those symptoms and treatments. Similarly, most of the individuals belonging to the cluster with longer LHS (40 days) had respiratory problems and loss of taste.

Considering the combination [LHS, S, T, Age], we obtained the clusters presented in Table 2. We want to point out that the same solution is also computed if the cluster classification is done in the cases [LHS, S, T, Age,

HT] and [LHS, S, T, PDP, HT]. We observe that for this combination of clusters, the variation of the age is the most important classifier, followed by LHS. They are the only two variables whose values varies across the clusters. This result is expected since the values of Age and LHS are larger than the others which are binary variables. So, the classification is not expected to be good. If the Age is weighted, we recover the first cluster combination.

Variables		Cluster 1	Cluster 2	Cluster 3	Cluster 4
	LHS	7	25	40	13
Symptoms	Muscular Pain	0	0	0	0
	Headache	0	0	0	0
	Fever	0	1	0	0
	Loss of Taste	0	0	1	0
	Loss of Smell	0	0	0	0
	Fatigue	0	1	0	0
	Respiratory	0	0	1	0
	Gastrointestinal	0	0	0	0
	Others	0	0	0	0
Treatments	Herferon	0	0	0	0
	Interferon	1	1	1	1
	Heberferon	0	0	0	0
	Chloroquine and Kaletra	0	0	1	0
	Antibiotics	0	0	0	0
	Others	0	1	0	0

Table 1: First cluster combination

Variables		Cluster 1	Cluster 2	Cluster 3	Cluster 4
LHS		14	12	12	11
Symptoms	Muscular Pain	0	0	0	0
	Headache	0	0	0	0
	Fever	0	0	0	0
	Loss of Taste	0	0	0	0
	Loss of Smell	0	0	0	0
	Fatigue	0	0	0	0
	Respiratory	0	0	0	0

	Gastrointestinal	0	0	0	0
	Others	0	0	0	0
Treatments	Herferon	0	0	0	0
	Interferon	1	1	1	1
	Heberferon	0	0	0	0
	Chloroquine and Kaletra	0	0	0	0
	Antibiotics	0	0	0	0
	Others	0	0	0	0
PDP	Age	58	44	79	23

Table 2: Second cluster combination

Finally, for the remaining combinations [LHS, S, T, PDP, H] and [LHS, S, T, Age, H], the use of K-means produced solutions which were like those obtained by the second combination. In fact, the values of LHS and the ages of the centroids were very similar, and the values of the other variables did not change across the clusters. Weighted distances did not improve the cluster classification because the first combination was recovered. So, in principle we have obtained two classifications, one mostly based on patients' age, and another based on symptoms and diseases.

Final Cluster Centres

Variables		Cluster 1	Cluster 2	Cluster 3	Cluster 4
	LHS	14	11	12	11
Symptoms	Muscular Pain	0	0	0	0
	Headache	0	0	0	0
	Fever	0	0	0	0
	Loss of Taste	0	0	0	0
	Loss of Smell	0	0	0	0
	Fatigue	0	0	0	0
	Respiratory	0	0	0	0
	Gastrointestinal	0	0	0	0
	Others	0	0	0	0
Treatments	Herferon	0	0	0	0
	Interferon	1	1	1	1
	Heberferon	0	0	0	0
	Chloroquine and Kaletra	0	0	0	0

	Antibiotics	0	0	0	0
	Others	0	0	0	0
PDP data	Edad	57	43	79	23
H data	DM	0	0	0	0
	Asthma	0	0	0	0
	HT	1	0	1	0
	Allergies	0	0	0	0
	Smoker	0	0	0	0
	Cancer	0	0	0	0
	IHD	0	0	0	0
	Others	0	0	0	0

Table 3: Third cluster combination

Now we will test the obtained classifications.

5.2. Classification of the individuals in the test sample

In this part we will demonstrate individuals' classification according to the three cluster families computed in the previous part. For each family f, f=1,2,3, we considered the combination $[LHS, C]_f$ that defined it. Then, centroid s_j=1,2,3,4 of the family f was given by the vector $[LHS_{fj}, C_{fj}]$, where LHS_{fj} is the LHS of the centroid j and C_{fj} is the value of the data included in the variables C, that defined family f. Similarly, for each individual I=1,2,.53 of the test sample, we took the partition of the data as $[LHS_I, C_I]$. Individual I will be assigned to the cluster k of family f, if

$$k \in arg\left\{\min_j \|C_I - C_{fj}\|\right\}$$

For testing the classification, we considered the value of LHS_i-LHS_{fk}. Individual I is a well-classified element if

$$|LHS_I - LHS_{fk}| < 3.$$

This means that the difference between the estimated value of LHS and the actual value is at most 2 days. From the viewpoint of the health logistics, this is not a large difference, i.e., the approximation is good.

Although in general the absolute value is important, we have also analysed in which cases LHS is underestimated. Indeed, in order to plan resources

such as beds, staff, respiratory aid equipment, etc., it is better that the estimated hospitalization time is larger than the actual value of LHS i.e., $LHS_I < LHS_{fk}$,.

Figure 4 shows the number of individuals for each possible value of $|LHS_I - LHS_{fk}|$ with the first family. As we observe, in most of the cases it is smaller than or equal to 2. Only in 17 cases this difference is larger than 3. Large values were observed, but in most cases the classification will estimate more days than the observed value. We want to point out that underestimation of LHS ($LHS_I > LHS_{fk}$) only appeared in 5 patients with 1, 5, 12 13 and 21 days less of LHS than estimated (one individual in each case).

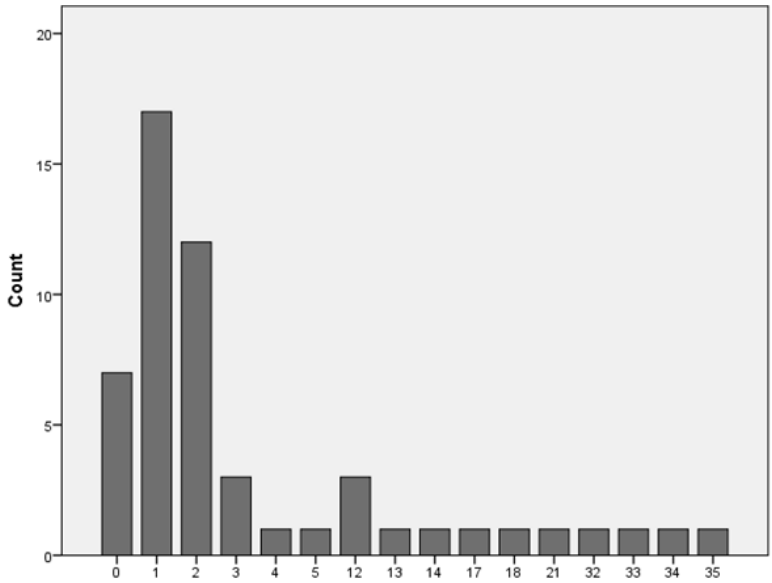


Figure 4. Combination 1

For the second family, the classification was worse. Figure 5 shows that in many cases (39 individuals), the absolute value of $LHS_I - LHS_{fk}$ is larger than 2. Underestimation was observed in 9 cases, with a difference of 1,2,2,3, 12,13,10,18,23 days.

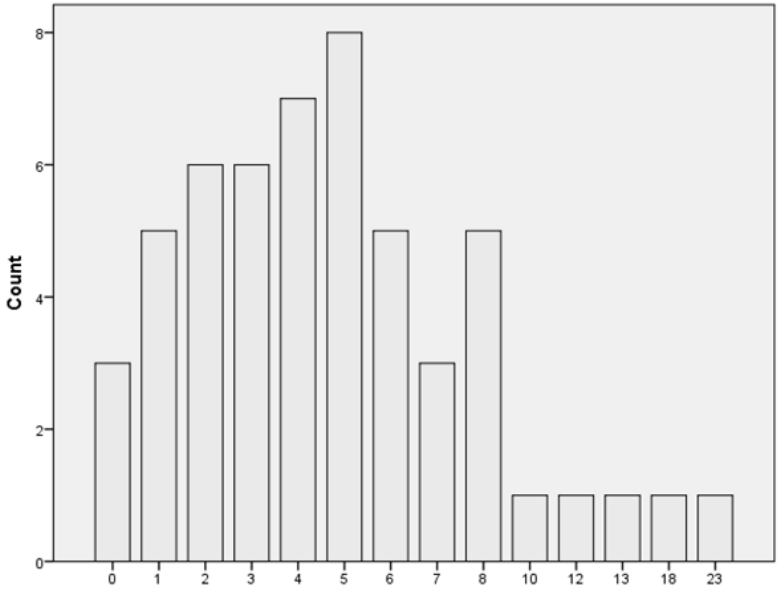


Figure 5. Combination 2

The third case is slightly better. However, see Figure 6, $|LHS_{app} - LHS| > 2$ for 37 individuals. Less days at the hospital were estimated in 9 cases, with two individuals with a difference of 2 days, and the other seven with an underestimation of 1, 3, 10, 12, 14, 19, 23 days.

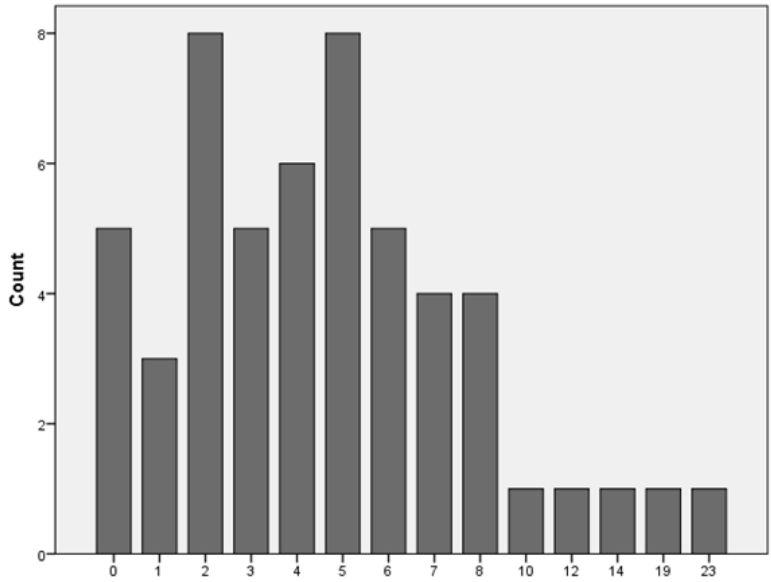


Figure 6: Combination 3

Taking the three combinations, we observe that the best classification was obtained by the first one. The following diagram shows in how many cases each combination was the best:

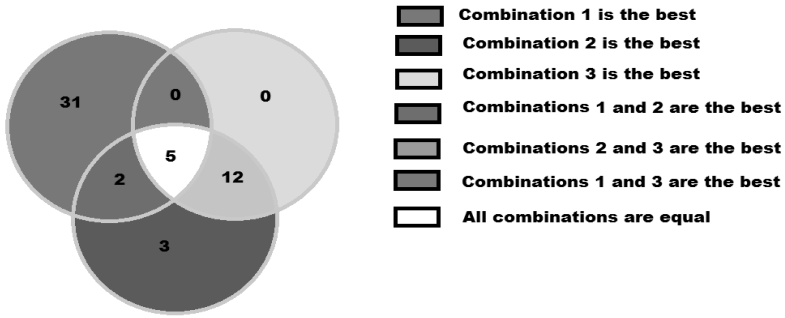


Figure 7. Best family

Compared with the results in [20], our approach provides a classification with less data. The LHS they estimated was more complex. The data included Urea and Haemoglobin, tests and a detailed description of the

medical treatments. With less data, we have obtained similar values of LHS. From the 16 values they forecasted, we have covered with an error, smaller than 3 days or less, 12 cases with less information. In the case of [30], the models predicted a stay of 8-9 days for patients without complications, which is similar to our estimation of 7 days. With complications, two models predicted 12-13 and another almost 19 days. We have provided more variability, since there were patients staying for more than a month.

6. Conclusions and remarks

In this work we have analysed data related to a sample of Cuban patients recovered from Covid-19. We performed a statistical analysis of the data. We obtained a positive significant correlation when comparing age with suffering from diabetes and HT, diabetes and experiencing fatigue and gastrointestinal symptoms and for having gastrointestinal symptoms and being treated with antibiotics. Cuban daily reports also highlight the prevalence of these factors in poor diagnosed patients, see [21] and [16]. The results were like those reported in the literature. As already mentioned in [10] and [29], a similar analysis was carried out. They included more data, such as oxygen saturation and blood tests, but the results of these tests were not available to the patient at the moment of the interview. So, we were able to obtain similar results with less data.

A more exhaustive analysis was developed with respect to the relation between LHS and the rest of the collected data. As expected, LHS is longer for older patients. Moreover, individuals older than 30 years stay on average 2.6 days more, and this difference is significant. Analogously, B+ patients stay 3.2 days longer than O+ individuals. The difference between the LHS of A+ people and the LHS of O+ and B+ is not significant. The large quantity of missing data for the variable blood type implies we could not analyse the other types of blood. If these data can be recovered, the study can be completed.

The analysis has concluded that diseases and symptoms had a significant correlation with suffering diabetes, having symptoms of fatigue, fever and gastrointestinal. These three symptoms are also significantly correlated with IHD.

Patients with IHD will, on average, stay longer at the hospital, and this difference is significant. Similarly, longer LHS are to be expected for patients having fatigue or fever. Fever was also reported by [14] as the main

symptom of COVID-19 patients and as very significant associated with a longer LHS. Although the p-value of the diabetes and gastrointestinal symptoms was larger than 0.05, it was close to this value. Hence, we need more data to clarify this result. As expected, the use of antibiotics, Chloroquine and Kaletra and other medicaments implies a longer LHS. This is expected by physicians, since they are prescribed when complications are present. The use of Herferon leads to the smallest average of LHS, but this difference is not significant. Note that this is not a concluding remark, since in many cases the information with respect to the kind of immunological booster they receive was very inaccurate.

Although the number of individuals was small, we could illustrate how using the procedure of the K-means for classifying the individuals and proposing an estimation of the LHS. The first obtained family was able to predict the 72% of the cases with an error of 2 days at the most. Only in the 9.4% of the cases, the value was underestimated. So, we can test this clustering with a more representative sample to check its robustness. Compared with [20], we were able to obtain a good classification with less data. This is an example of the use of classification techniques to health problems. Other instances can be found in [27].

The main drawback of this classification approach is its sensitivity to changes in the characteristics of the diseases. For instance, if a new strain appears, LHS can vary dramatically. In this case, the analysis must be repeated. Note that Delta strain was not present in the country when the sample was collected.

Further data analysis can be done with other data that have been collected afterwards. For instance, for each patient, it is known the number of contacts they have declared and how many of them got the disease. Estimating the number of contagious, given the contacts, is another interesting idea. This can be developed by differentiating among the members of the family, co-workers and other groups. Similarly, relations among symptoms, treatments, LHS, PDP, D and H and the diseases that can appear after the infection can be also analysed for Cuban patients and compared with the findings in [15].

Present study points to LHS as a remarkable element to face COVID-19 pandemic. Resource administration is needed to avoid health systems collapse and to predict which patients can be at risk of dying/having a prolonged LHS according to COVID-19 clinical featuring (as we did here)

and lab biomarkers/ prognosis scales. A future, larger study, with a more representative sample, is needed to assure to the associations we have been successful in reaching the most of the Cuban/Havana's population. New data will be gathered, so this preliminary study could be completed anytime.

Acknowledgments

This research has been developed as part of the project PN223LH010-005 Desarrollo de nuevos modelos y métodos matemáticos para la toma de decisiones, supported by Oficina de Gestión de Fondos y Proyectos Internacionales del Ministerio de Ciencia, Tecnología y Medio Ambiente de la República de Cuba.

References

AL-Gahtani, S. and Shoukri, M.M. (2021): Analysis of Length of Stay (LOS) Data from the Medical Records of Tertiary Care Hospital in Saudi Arabia for Five Diagnosis Related Groups: Application of Cox Prediction Model. *Open Journal of Statistics,* 11: 99-112. https://doi.org/10.4236/ojs.2021.111005.

Allende, S, Bouza, G and Sariol, W. (2020): COVID-19 en Cuba: Estimación del número de camas diarias que se requieren durante un período de la epidemia, *Ciencias Matemáticas*, 34(1):105-111.

Atkins J.L, Masoli JA, Delgado J, Phillling L, Kuo CL, Kuchel GA. (2020): Pre-existing comorbidities predicting COVID-19 and mortality in the UK Biobank Community Cohort. *J Gerontol* A *Bio Sci Med Sci*, 75: 2224- 2230.

Aunan JR, Cho WC, Søreide K. (2017): The Biology of Aging and Cancer: A Brief Overview of Shared and Divergent Molecular Hallmarks. *Aging Dis*. 2017;8(5):628-642. Oct 1. doi:10.14336/AD.2017.0103

Bondeelle L, Chevret S, Cassonnet S, Harel S, Denis B, de Castro N, et al. (2021) Profiles and outcomes in patients with COVID-19 admitted to wards of a French Oncohematological hospital: Aclustering approach. *PLoS ONE* 16(5): e0250569. https://doi.org/10.1371/journal.pone.0250569).

Burke GM, Genuardi M, Shappell H, D'Agostino RB Sr, Magnani JW. (2017): Temporal Associations Between Smoking and Cardiovascular Disease, 1971 to 2006 (from the Framingham Heart Study). *Am J Cardiol*.120(10):1787-1791. doi:10.1016/j.amjcard.2017.07.087.

Cazzoletti L, Ferrari M, Olivieri M, Verlato G, Antonicelli L, Bono R, Casali L, Cerveri I, Marchetti P, Pirina P, Rossi A, Villani S, de Marco

R. (2015): The gender, age and risk factor distribution differs in self-reported allergic and non-allergic rhinitis: a cross-sectional population-based study. *Allergy Asthma Clin Immunol.* Dec 4; 11:36. doi: 10.1186/s13223-015-0101-1. PMID: 26640494; PMCID: PMC4669616.

Chen F-J, Li F-R, Zheng J-Z, Zhou R, Liu H-M, Wu K-Y, et al. (2021): Factors associated with duration of hospital stay and complications in patients with COVID-19 *J Public Health Emerg.*; 5:6 http://dx.doi.org/10.21037/jphe-20-74).

Chien YW, Wang CC, Wang YP, Lee CY, Perng GC. (2020): Risk of Leukemia after Dengue Virus Infection: A Population-Based Cohort Study. *Cancer Epidemiol. Biomarkers Prev*.;29(3):558-564. doi: 10.1158/1055-9965.EPI-19-1214.

Fang H, Liu Q, Xi M, Di X, He J, Luo P, et al. (2021): Impact of comorbidities on clinical prognosis in 1280 patients with different types of COVID-19. *J. Investig. Med*.: 69: 75-85.

Forsyth, David (2018): Statistical inference as severe testing. How to get beyond the Probability and Statistics for Computer Science, Springer International Publishing ISBN978-3-319-64409-7.

Gentle J.E., W.K. Hardle and Y. Mori, Editors. (2012): Handbook of Computational Statistics. Concepts and methods. (2ND ED. REVISED AND UPDATED) Springer, ISBN 978-3-642-21550-6.

Giordani, P,Ferraro, MB,Martella, F. (2020)*An Introduction to Clustering with R*, Springer Verlag.

Guo A, Lu J, Tan H, Kuang Z, Luo J, Yang T et al. (2021): Risk factors on admission associated with hospital length of stay in patients with COVID-19:a retrospective cohort study. *Scientific Reports* 11:7310.

Huang C, Huang L, Wang Y, Li X, Ren L, Gu X, et al. (2021): 6- month consequences of COVID -19 in patients discharged from hospital: a cohort study. *Lancet*; 397: 220-232.

León Álvarez Jorge Luis, Calderón Martínez Marcy, Gutiérrez Rojas Angela Rosa (2021): Análisis de mortalidad y comorbilidad por COVID-19 en Cuba. *Rev Cubana Med. [online]*. in: http://scielo.sld.cu/scielo.php?script=sci_arttext&pid=S0034-75232021000200004&lng=es.

Li G, Liu Y, Jing X, Wang Y, Miao M, Tao L. (2021): Mortality risk of COVID-19 in elderly males with comorbidities: a multi-country study. *Aging*;13(1): 27- 60.

Liu, M.; Cao, J.; Liang, J.; Chen, M. J. (2020): *Epidemic-logistics modelling: a new perspective on Operations Research*. Springer.

Magen E, Yosefy C, Viskoper R, Mishal J.(2006): Treatment of allergic rhinitis can improve blood pressure control. *J. Hum. Hypertens.* 20**,** 888–893. https://doi.org/10.1038/sj.jhh.1002088.

Mahboub B, Bataineh MTA, Alshraideh H, Hamoudi R, Salameh L and Shamayleh A (2021) Prediction of COVID-19 hospital length of stay and risk of death using Artificial Intelligence-based modelling. Front. Med. 8:592336. doi: 10.3389/fmed.2021.592336).

Ministerio de Salud Pública de Cuba*, https://salud.msp.gob.cu/.*

Ministerio de Salud Pública de Cuba. Dirección de Registros Médicos y Estadísticas de Salud. *Anuario Estadístico de Salud 2020.* La Habana: Ministerio de Salud Pública; 2021. ISSN: 1561-4433.

Rodríguez-Guzmán R, Céspedes EM, Guzmán- Díaz P. (2021): Endothelial dysfunction and cardiovascular diseases through oxidative stress pathways. In *Shampa C. Endothelial Signalling in Vascular Dysfunction and Disease*. Academic Press. ELSEVIER. Chapter 19: 213- 219. ISBN: 978-012-816196-8. https://www.elsevier.com/books/endothelial-signaling-in-vascular-dysfunction-and-disease/chatterjee/978-0-12-816196-8).

Rubio MC, Sanchez L, Abreu-Ruíz G, Bermejo-Bencomo W, Crombet T, Lage A. (2020): COVID-19 and Cancer in Cuba. *Semin. Oncol*.;47(5):328-329.

Scioli MG, Storti G, D'Amico F, Rodríguez Guzmán R, Centofanti F, Doldo E, Céspedes Miranda EM, Orlandi (2020): A. Oxidative Stress and New Pathogenetic Mechanisms in Endothelial Dysfunction: Potential Diagnostic Biomarkers and Therapeutic Targets*. J Clin Med*. Jun 25;9(6):1995. doi: 10.3390/jcm9061995. PMID: 32630452; PMCID: PMC7355625.

Ssentongo P, Ssentongo AE, Heilbrunn ES, Ba DM, Chinchilli VM. (2020): Association of cardiovascular disease and 10 other pre-existing comorbidities with COVID-19 mortality: a systematic review and meta-analysis. *PLoS ONE*.; 15(8): e0238215.

Subasi, A. (2020): Practical Machine Learning for Data Analysis Using Python *Elsevier*.

Thakur B, Dubey P, Benitez J, Torres JP, Reddy S, Shokar N, et al. (2021): A systematic review and meta-analysis of geographic differences in comorbidities and associated severity and mortality among individuals with COVID-19. *Scientific Reports*.; 11: 8562.

Thiruvengadam G, Lakshani M, Ramanujam R. (2021): A study of factors affecting the length of hospital stay of COVID-19patients by Cox-Proportional Hazard model in a South Indian Tertiary Care Hospital *Journal of Primary Care & Community Health*, 12: 1-7

https://doi.org/10.1186/s12879-021-06371-6.
Vekaria B, Overton C, Wisniowski A, Ahmad S, Aparicio-Castro A, Curran-Sebastian J *et al.* 2021): Hospital length of stay for COVID-19 patients: Data-driven methods for forward planning, *BMC Infectious Diseases* 21: 700.
Wu, J. (2012): Advances in K-means Clustering, A data mining thinking, Springer Berlin Heidelberg.
Zhou F, Yu T, Du R, Fan G, Liu Y, Liu Z, et al. (2020): Clinical course and risk factors for mortality of adult patients with COVID-19 in Wuhan, China: a retrospective cohort study. *Lancet;* 395: 1054-62.

CHAPTER IX

PREDICTIVE VARIABLES OF LUNG DAMAGE IN RECOVERED COVID-19 PATIENTS

CARMEN VIADA GONZÁLEZ[1],
PATRICIA LORENZO LUACES[1],
LISANIA REYES ESPINOSA[1],
AGUSTÍN LAGE DÁVILA[1],
CONSUELO MACÍAS ABRAHAM[2], ANA MARÍA,
SIMÓN PITA[2], LAURA RUIZ VILLEGAS[2]
AND AMPARO MACÍAS ABRAHAM[1]

1. Introduction

The 2019 novel coronavirus (COVID-19) pandemic, which causes severe acute respiratory syndrome (SARS-CoV-2, coronavirus 2), has affected more than 61 million people worldwide, since its beginning in Wuhan in December 2019 [WHO, Covid-19]. This disease is characterized by severe pneumonia with fever, respiratory symptoms, and pulmonary lesions in the form of ground glass opacity, detected on chest CT [Li, 2020]. From the immune-haematological point of view, lung hyper inflammation has been evidenced at the expense of neutrophils and pro-inflammatory cytokines (IL-6, IL-10) [Zheng, 2020; Liu J, 2020]. Significant lymphopenia has also been described (CD4 + T and CD8 + T, and in some cases B cells) and decrease in NK cells, whose kinetics correlate with the degree of severity of the patients [Wang, 2020; Wan, 2020]. In addition to a state of hypercoagulability given by increased D-dimer levels, prolonged

[1] Centre of Molecular Immunology, Cuba
[2] Institute of Hematology and Immunology

prothrombin time, and low platelets, leading to pulmonary micro-thrombosis [Salamanna, 2020; Liu Y, 2020]. The acute phase of the disease has been widely described; however, there are few studies in the recovery phase.

In Cuba, more than 90% of patients have recovered from the disease, but one third of them have been left with inflammatory lung damage that can lead to irreversible pulmonary fibrosis. For this reason, a clinical trial has been conducted to evaluate treatment with autologous stem cells or prednisone in recovered patients who were left with lung damage, in order to prevent pulmonary fibrosis, the results of which will be published shortly. Based on this trial, it was interesting to identify inflammation markers that allow predicting lung damage in recovered COVID-19 patients and identifying individuals at risk, in whom we must intervene to prevent or minimize damage. Inflammation indexes have been reported to be inflammation markers easy to identify, studied in different types of cancer [Lolli, 2016; Howard, 2019], infections [Takada, 2019], autoimmune and inflammatory diseases [Kucuk, 2020; Çetinkaya, 2020], COPD [Paliogiannis, 2018], diabetes mellitus [Palella, 2020], and more recently in COVID-19 [Liu J, 2020]. We set out to evaluate the behaviour of inflammation indexes in recovered COVID-19 patients with or without lung damage; to determine the best index to predict lung damage; and to evaluate the relationship of these indexes with the treatment of patients with lung damage.

2. Patients and Methods

A prospective longitudinal exploratory study has been conducted in 49 recovered COVID-19 patients attended in the Institute of Haematology and Immunology in collaboration with the Centre of Molecular Immunology, 27 with lung damage and 22 without damage. The study has been approved by the research and ethics committee and the Cuban regulatory agency. The 20 patients with damage came from a non-randomized, controlled clinical trial in order to treat these coronavirus sequelae through regenerative autologous stem cell therapy or prednisone. Autologous stem cells were administered to 10 patients by intravenous infusion in a single dose of 200 x 10^6 mononuclear cells obtained from peripheral blood with at least 1% CD34 + progenitor cells. Stem cells were previously mobilized with the granulocyte-colony stimulating factor (G-CSF) (IOR®LEUKOCIM) at a dose of 40 µg / kg and prednisone therapy was administered at 0.75 mg / kg per day to the other 10 patients. The results of this clinical trial will be published shortly.

The twenty-nine recovered patients who did not have lung damage were not part of the clinical trial, but they were included in the present study.

The 49 patients underwent a complete blood count, obtaining the absolute count of neutrophils (N) and lymphocytes (L). In the analyses, the neutrophil lymphocyte ratio (NLR) was defined as NLR=N/L, platelet lymphocyte ratio (PLR) was defined as PLR=P/L and systemic immune-inflammation index (SII) was defined as SII=(N*P)/L. Inflammation indexes were compared in recovered patients with lung damage and without lung damage. In patients with lung damage, the markers were compared between the group that received autologous stem cells and the group treated with prednisone, at the initial time and at one month of treatment; and finally, within each treatment group, the indexes were compared at the initial moment and at one month of treatment. In addition, it was analysed which index was better to predict lung damage, according to discrimination capacity.

3. Statistical analysis

Kolmogorov-Smirnov test was performed to determine normality of the variables, non-parametric comparisons Mann-Whitney U or t-student test according to the distribution of the analysed variables. Receiver Operating Characteristic (ROC) curves were analysed to determine the optimal cut-off for classification of lung damage and it estimated the sensibility, specificity and area under curve. To determine the importance of the three indexes in the classification of lung damage, binary logistic regression was applied, taking lung damage as the dependent variable, and it estimated the relative risk (RR). All statistical analyses were carried out using IBM SPSS Statistics Version 25 software. Statistically significant differences were considered when $p < 0.05$.

4. Results and discussion

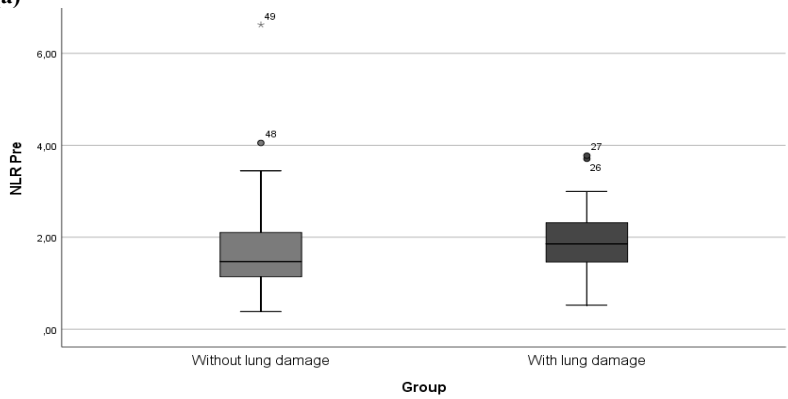
A total of 49 patients recovered from COVID-19 were studied, with a median age of 52 years (from 21 to 75 years), 53.1% of them were female and 65.3% had mild forms of the disease (without requirement of oxygen therapy) during the active stage of the disease. Among patients include, 55.1% who had mild forms of the disease did have lung damage during recovery phase. Of the patients with lung damage, 65% had been reported as severe or critical during the acute phase.

Variable	N (%)
Age	
Mean (Min, Max)	52 (21,75)
Sex	
Female	26 (53,1)
Male	23 (46,9)
Oxygen therapy	
With **requirement**	17 (34,7)
Without **requirement**	32 (65,3)
Lung damage	
With damage	27 (55,1)
Without damage	22 (44,9)
With lung damage	
Severe or critical	17 (65,0)
Care	10 (35,0)

Table 9.1. Characteristics of the patients

To analyse each of the three inflammation indexes and their relationship with lung damage, the behaviour of the NLR, PLR and SII in patients with lung damage was compared with those without damage. In all three cases, the markers were higher in patients with damage than in those without sequelae, which could suggest that even after recovery, an important inflammatory component that causes lung damages persists. This difference was statistically significant in PLR and SII. Levels of PLR and SII in patients without or with lung damage are shown in (Figure 1). In the case of the NLR, there were no significant differences between the two groups; however, patients with lung damage had a higher NLR.

(a)



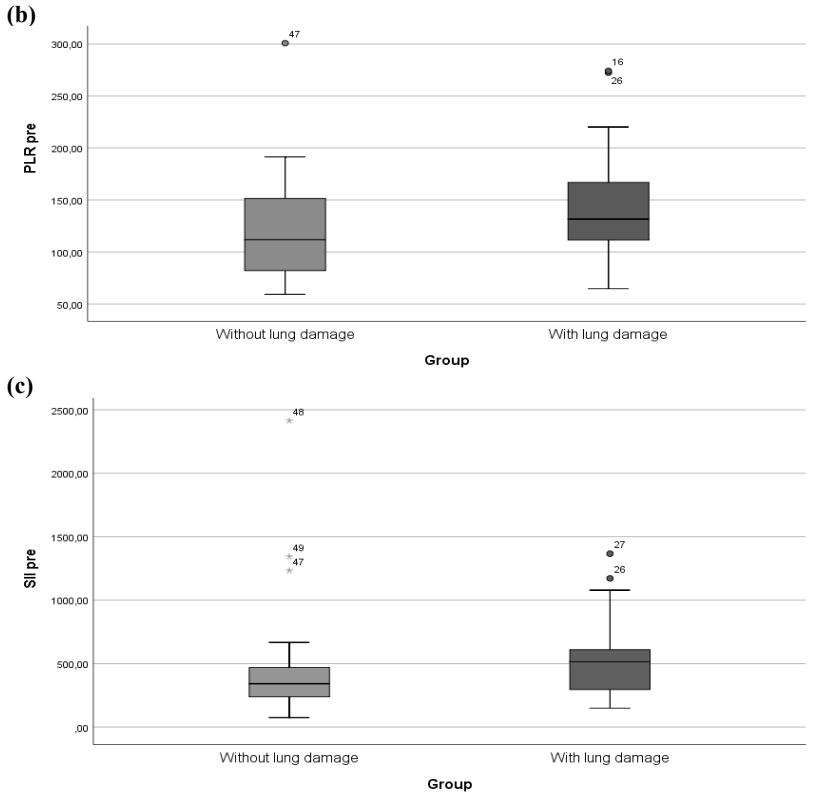


Figure 9.1 Levels of NLR (a) PLR (b) and SII (c) in patients without or with lung damage

Box plot representing the baseline levels of NLR (a), PLR (b) and SII (c) in patients without or with lung damage. Student's t-test was performed for PLR analysis according to its normal distribution, $^{*}p = 0.031$. U-Mann Whitney test was performed for SII analysis according to its non-normal distribution, $^{*}p = 0.02$.

To identify which of the three indexes constituted a predictive marker of lung damage, they were analysed in two ways: as continuous variables and as dichotomous variables.

In the approach of continuous variables, the selected variable was PLR, with 67.3% of good classification: it correctly selected 89.7% of those without

lung damage, but only 35% of those with lung damage. The inclusion of the other two indexes did not increase the classification of patients.

For the analysis of the indexes as dichotomous variables, ROC curve analysis was first applied to find cutoff for NLR, PLR and SII that can differentiate patients with or without lung damage (Figure 9.2, Table 9.2).

Inflammation index	Cutoff	Sensibility	Specificity	Area under curve
NLR	1.54	0.85	0.552	0.640
PLR	112.07	0.85	0.552	0.702
SII	340.82	0.8	0.552	0.698

Table 9.2 Cutoff determined for each index

Cutoff, sensitivity, specificity, and area under curve for NLR, PLR and SII are shown. ROC curves were applied.

Dichotomizing the three indexes according to the cutoff points, a statistically significant relationship was obtained between the high indexes (above the cutoff) and the presence of lung damage, which suggests that the higher the index, the greater the possibility of having lung damage (Table 9.3).

Index	With lung damage (%)	Without lung damage (%)	Total (%)	p-valor
NLR >=1.54**	17 (85,0)	13 (44,8)	30 (61,2)	0.005
PLR >=112.07**	17 (85,0)	13 (44,8)	30 (61,2)	0.005
SII >=340.82*	16 (80,0)	13 (44,8)	29 (59,2)	0.014

Table 9.3 Association of the indexes with lung damage

Number of patients with high levels of inflammation indexes (above the cutoff), separated into two groups with or without lung damage and the total number of patients. The data is shown in number and percent. Binary logistic regression was applied, taking lung damage as the dependent variable, $^{**}p = 0.005$, $^{*}p = 0.014$

When the predictive capacity of the damage was analysed, in the case of the PLR, a good classification of 67.3% was obtained as a dichotomous variable. Among patients included, 85% who had lung damage were well-classified and only 55.2% of those did not. PLR levels higher than 112.07 were associated with an increase of almost 7 times the risk (Table 9.4).

Observed	**Predicted**		
	With lung damage	**Without lung damage**	**Percentage Correct**
Without lung damage	13	16	55,2
with lung damage	17	3	85,0
Overall Percentage			67,3
RR 6,974			

Table 9.4. PLR as a dichotomous variable to predict lung damage

The observed and predictive number of patients with and without lung damage, the percentage of correct classification, the global percentage and the relative risk (RR) are shown. Binary logistic regression was applied, taking lung damage as the dependent variable.

If PLR and NLR are included, a 73.5% good rating was obtained. 80% of the patients who had lung damage and 69% of those who did not have it are correctly selected. The risk of damage in a patient with PLR value greater than 112.07 and NLR greater than 1.54 is almost 4 times greater than a patient with a lower PLR.

The SII did not score well on its own, or in combination with either of the other two indexes.

To study the effect of the treatment on the inflammation of the patients with lung damage, each index of inflammation was analysed after one month of treatment, compared to the initial moment, when the patients had not yet been treated. The PLR decreased in patients one month after being treated with prednisone, an effect that was not statistically significant in the group treated with autologous stem cells. PLR levels before and one month after prednisone shown in (Figure 9.2).

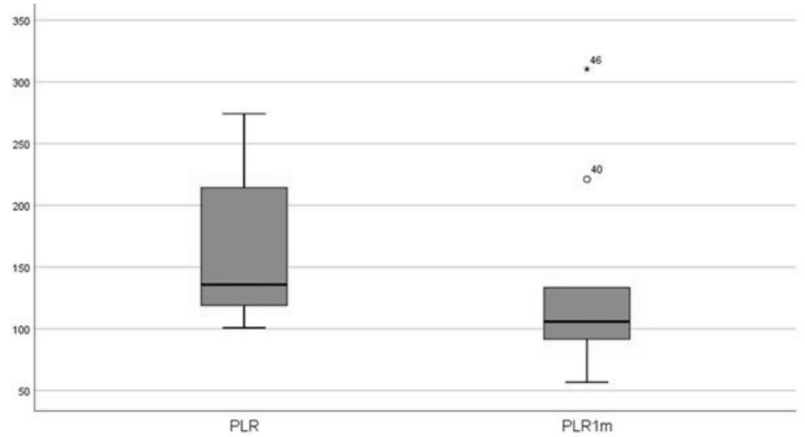


Figure 9.2 PLR levels one month after treatment with prednisone

Box plot representing PLR levels prior to treatment of patients (PLR) and one month after treatment with prednisone (PLR1m). Wilcoxon test was performed, $^{*}p = 0.047$.

The NLR and SII did not show significant differences at one month of treatment neither with autologous stem cells nor with prednisone (Table 9.5).

	Autologous stem cells		**Prednisone**	
	Mean (CI95%)	**Median**	**Mean (CI95%)**	**Median**
NLR	2.1 (1.5; 2.7)	2.0	2 (1.5; 2.6)	1.8
NLR1m	1.8 (1.0; 2.5)	1.3	2.7 (1.9; 3.6)	2.8
PLR	147.4 (116.1; 178.8)	150.9	166.3 (119.5; 213.1)	135.8
PLR1m	136.7 (95.5; 178)	128.6	133.1 (78.7; 187.6)	105.8
SII	610 (374.9; 845.1)	550.2	541.1 (343.3; 738.9)	521.7
SII1m	516.5 (284.6; 748.4)	400.3	674.8 (425.6; 923.9)	645.3

Table 9.5 Behaviour of the three indexes one month after the treatment with respect to the initial moment

Mean, confidence interval (CI) and median of the NLR, PLR and SII before the start of treatment and one month after treatment (1m) with autologous stem cells or prednisone are shown. A Wilcoxon test was performed on the three indexes.

5. Conclusions

The baseline levels of PLR and SII were different in patients without or with lung damage.

The cutoff points found for the inflammation indices had high sensitivity, specificity and the area under the curve.

PLR and SII are predictive markers of lung damage, with PLR being the one that best discriminates between damaged patients.

These results suggest that prednisone treatment could be used in recovered patients with lung damage and a PLR greater than or equal to 112.07.

Bibliography

Çetinkaya ÖA, Çelik SU, Terzioğlu SG et al. The Predictive Value of the Neutrophil-to-Lymphocyte and Platelet-to Lymphocyte Ratio in Patients with Recurrent Idiopathic Granulomatous Mastitis. Eur J Breast Health. 2020;16(1): 61-65. Howard R, Kanetsky PA, Egan KM. Exploring the prognostic value of the neutrophil-to-lymphocyte ratio in cancer. Sci Rep. 2019;9(1):19673.

Kucuk H, Tecer D, Goker B et al. Platelet/lymphocyte ratio and mean platelet volume in patients with granulomatosis with polyangiitis. Adv Rheumatol. 2020; 60(1): 4.

Li K, Wu J, Wu F et al. The Clinical and Chest CT Features Associated with Severe and Critical COVID-19 Pneumonia. Invest Radiol. 2020; 55(6): 327-331.

Liu J, Li S, Liu J et al. Longitudinal characteristics of lymphocyte responses and cytokine profiles in the peripheral blood of SARS-CoV-2 infected patients. EBioMedicine. 2020:55:102763.

Liu J, Liu Y, Xiang P et al. Neutrophil-to-Lymphocyte Ratio Predicts Severe Illness Patients with 2019 Novel Coronavirus in the Early Stage. J Transl Med. 2020; 18(1): 206.

Liu Y, Sun W, Guo Y et al. Association between platelet parameters and mortality in coronavirus disease 2019: Retrospective cohort study. Platelets. 2020; 31(4): 490-496.

Lolli C, Caffo O, Scarpi E et al. Systemic Immune-Inflammation Index Predicts the Clinical Outcome in Patients with mCRPC Treated with Abiraterone. Front. Pharmacol. 2016; 7: 376.

Palella E, Cimino R, Pullano SA et al. Laboratory Parameters of Hemostasis, Adhesion Molecules, and Inflammation in Type 2 Diabetes Mellitus: Correlation with Glycemic Control. Int. J. Environ. Res. Public Health. 2020; 17(1): 300.

Paliogiannis P, Fois AG, Sotgia S et al. Neutrophil to lymphocyte ratio and clinical outcomes in COPD: recent evidence and future perspectives. Eur Respir Rev. 2018; 27(147): 170113.

Salamanna F, Maglio M, Landini MP et al. Platelet functions and activities as potential hematologic parameters related to Coronavirus Disease 2019 (Covid-19). Platelets. 2020; 31(5): 627-632.

Takada T et al. Added value of inflammatory markers to vital signs to predict mortality in patients suspected of severe infection. Am J Emerg Med. 2019; 38(7): 1389-1395.

Wan S, Yi Q, Fan S et al. Relationships among Lymphocyte Subsets, Cytokines, and the Pulmonary Inflammation Index in Coronavirus (COVID-19) Infected Patients. Br J Haematol. 2020; 189: 428–437.

Wang F, Nie J, Wang H et al. Characteristics of Peripheral Lymphocyte Subset Alteration in COVID-19 Pneumonia. JID. 2020;221(11):1762-1769.

World Health Organization. COVID-19 Weekly Epidemiological Update. https://www.who.int/publications/m/item/weekly-epidemiological-update---1-december-2020. Accessed 2 Dec 2020.

Zheng M, Gao Y, Wang G et al. Functional exhaustion of antiviral lymphocytes in COVID-19 patients. Cell Mol Immunol. 2020;17(5):533-535.