Correlational analysis between the economics, socio-demographic indices and statistics of contagion due to Covid-19, applying the Clustering methodology in countries of America

Análisis correlacional entre la economía, los índices sociodemográficos y las estadísticas de contagio por Covid-19, aplicando la metodología de Clustering en países de América

DOI: http://doi.org/10.17981/ingecuc.17.1.2021.21

Artículo de Investigación Científica. Fecha de Recepción: 01/12/2020. Fecha de Aceptación: 25/01/2021.

Elisa del Carmen Navarro-Romero
Universidad Santo Tomás. Bogotá, D.C. (Colombia)
elisanavarro@usantotomas.edu.co

Óscar Mauricio Gelves-Alarcón
Universidad Militar Nueva Granada. Bogotá, D.C. (Colombia)
oscar.gelves@unimilitar.edu.co

Natalia García-Corrales
Environment & Technology Foundation. Cereté (Colombia)
etf@environmenttechnologyfoundation.org

To cite this article:
E. Navarro Romero, O. Gelves Alarcón & N. Carcía Corrales, " Correlational analysis between the economics, socio-demographic indices and statistics of contagion due to Covid-19, applying the Clustering methodology in countries of America", *INGECUC*, vol. 17. no. 1, pp. 285–302. DOI: http://doi.org/10.17981/ingecuc.17.1.2021.21

Abstract

Introduction— This research is motivated, by the current world situation, caused by the pandemic declared by the WHO before the spread and severity of the coronavirus disease (COVID-19), notified for the first time in Wuhan (China) on December 31 of 2019. Through mathematical and statistical analysis, it seeks to show and explain in an expeditious manner, the causes for which there is a higher rate of contagion and lethality due to the virus, in different countries, taking into consideration patterns associated with social political behavior and economic, as a first approach to knowing statistics that allow generating forecasts for future periods, given the conditions.

Objective— The main objective of this work is to define the correlation of the economic, social and demographic variables of the countries of America with respect to the contagion of the virus, proposing a forecast model on the level of contagion in each cluster proposed by the different regions of the American continent.

Methodology— The study performs clustering (grouping) of the countries of America with respect to their geographical position North America, Central America and the Caribbean islands and South America, followed by a search for statistical data on social, economic and demographic indicators of the countries of America in recent years and statistics of levels of contagion of COVID 19 in sources such as international organizations regulating health issues. Next, a characterization and correlation of the collected data was carried out, to finally, based on the results of the correlation, make a forecast of the level of contagion that would be reached by each of the regions.

Results— The purpose of this document is to provide information on the countries of North America, Latin America and the Caribbean with respect to the analysis of mortality from COVID-19, through methods of analysis of mortality from all causes as one of the approaches proposed to contribute to the assessment of the true magnitude of the burden of the COVID-19 epidemic in these countries.

Conclusions— The results show interesting information, since the Latin American curve turned out to be much less pronounced than that of the United States, in terms of contagion and deaths, despite the socio-demographic conditions, economic, technological and political opportunities. This analysis invites us to find out which are those correlations that directly impact the behavior of infections, taking into account variables such as age, gender, stratum, level of education, and other sociodemographic characteristics that may influence the spread or containment of the virus.

Keywords— Grouping; data; ICR; mathematical model; pandemic; linear regression; multiple regression

Resumen

Introducción— Esta investigación está motivada, por la actual situación mundial, provocada por la pandemia declarada por la OMS ante la propagación y gravedad de la enfermedad por coronavirus (COVID-19), notificada por primera vez en Wuhan (China) el 31 de diciembre de 2019. A través del análisis matemático y estadístico, se busca mostrar y explicar de manera expedita, las causas por las cuales existe una mayor tasa de contagio y letalidad por el virus, en diferentes países, tomando en consideración patrones asociados al comportamiento político social y económico, como una primera aproximación para conocer estadísticas que permitan generar pronósticos para periodos futuros, dadas las condiciones.

Objetivo— El objetivo principal de este trabajo es definir la correlación de las variables económicas, sociales y demográficas de los países de América, con respecto al contagio del virus, proponiendo un modelo de pronóstico sobre el nivel de contagio en cada cluster propuesto por las diferentes regiones del continente americano.

Metodología— El estudio realiza una clusterización (agrupación) de los países de América con respecto a su posición geográfica América del Norte, América Central e islas del Caribe y América del Sur, seguido de una búsqueda de datos estadísticos sobre indicadores sociales, económicos y demográficos de los países de América en los últimos años y estadísticas de niveles de contagio del COVID 19 en fuentes como los organismos internacionales que regulan los temas de salud. Luego, se realizó una caracterización y correlación de los datos recolectados, para finalmente, en base a los resultados de la correlación, realizar un pronóstico del nivel de contagio que alcanzaría cada una de las regiones.

Resultados— El propósito de este documento es proporcionar información sobre los países de América del Norte, América Latina y el Caribe con respecto al análisis de la mortalidad por COVID-19, a través de métodos de análisis de la mortalidad por todas las causas como uno de los enfoques propuestos para contribuir a la evaluación de la verdadera magnitud de la carga de la epidemia de COVID-19 en estos países.

Conclusiones— Los resultados muestran información interesante, ya que la curva latinoamericana resultó ser mucho menos pronunciada que la de Estados Unidos, en términos de contagio y muertes, a pesar de las condiciones sociodemográficas, económicas, tecnológicas y políticas. Este análisis invita a averiguar cuáles son las correlaciones que impactan directamente en el comportamiento de los contagios, teniendo en cuenta variables como la edad, el género, el estrato, el nivel de educación y otras características sociodemográficas que pueden influir en la propagación o contención del virus.

Palabras clave— Agrupación; datos; ICR; modelo matemático; pandemia; regresión lineal; regresión múltiple

INGE CUC vol. 17 no. 1, pp. 285-302. Enero - Junio, 2021
Barranquilla. ISSN 0122-6517 Impreso, ISSN 2382-4700 Online

I. Introduction

Humanity has been affected by different pandemics throughout its history. This term alludes to diseases that attach themselves to individuals from diverse territories or multiple countries [1]. As stated by several researchers [2] the efficiency of a pandemic peaks with a person-to-person transmission. Within the pandemics it's worth mentioning the ones caused by virus such as H1N1, VIH, and most recently, SARS-CoV-2 (Coivid-19). The propagation of disease due to Coronavirus (COVID-19) has presented a menace to health around the globe. By March 17th of 2020 the number of confirmed cases had risen above 179 000, with more than 7 000 confirmed dead in at least 150 countries [3].

Granted, Covid-19 is considered an airborne transmitted disease, considering that the virus can transfer itself through sneezing, coughing, talking or any activity that result in the generation of aerosolized particles [4]. This requires establishing means of distancing people and generating protection mechanisms such as the use of face masks in environments where the virus can be transmitted with ease.

The impact to the populational density, regarding infectious diseases has been subject to various studies. Density referring to contact and interaction between inhabitants, which makes this variable crucial for the propagation of emerging infectious diseases. In the case of global pandemics such as the recent COVID-19 virus [5], the larger and denser urban centers, particularly those involved with tourist activities, could become in epicenters of a health crisis resulting in the death of thousands. In a similar manner, health & educational systems could help minimize the disease's total impact for those who have been infected, which leads to a higher recovery rate and a lower mortality rate. Densities tend to vary higher than the confirmed infection rates. The simple correlation between the two figures approximates to 0.48, which means that, for a country, the density only represents around 23% variation of the virus' propagation. Density and exposure are not a straight line. Dense areas can have a higher chance of putting into practice policies that ensure distancing, reducing real infection rates or simply leading to a large-scale social distancing due to a greater quantity of people conscious about the threat.

During this period the pandemic is taking place in, humanity has understood the importance of possessing data that allows processing, extracting information, and thus, predicting its behavior. As a number of studies [6] epidemiology and biostatistics oversee analyzing the patterns described within the population's diseases, and in mathematical terms, projecting their future. The Pan-American organization of health [7] highlights the importance of possessing high quality, accessible, trustworthy, opportune, open and reliable data analyzed recently that allow us to generate information geared towards decision making during a pandemic; while at the same time warning us about the difficulties caused by "infopedia", referring to the health problems rooted in the excess of unreliable information regarding the pandemic's current behavior, which resulted in trouble for Covid-19 related knowledge by February 2020 [8]. The concern the excess or lack of trustworthy information geared towards decision making that determines the population's behavior, and thus, influences their health, is approached as an analysis that provides a tool for measuring the concern of catching Covid-19, In a sample of the Peruvian population; which concludes, information is one of the factors that influences the behavior of a pandemic throughout a territory.

In accordance with different research [10], the metropolitan population is one of the most important infection predictors, populations with larger metropolitan areas, suffer higher infection indices and mortality rates. However, the opposite is true in regard to the internal level of the states or departments of the different countries, and the mortality rate, possibly due to bigger adherence to the politics and practices of social distancing and better health services. These findings suggest that connectivity is a greater factor towards Covid-19 propagation than density.

Furthermore, other research [11] a study was conducted regarding the correlation between the demographic variables and the lethality of Covid-19 in OCDE countries, considering them homogeneous. Result show that the mortality and lethality of Covid-19 are not associated with demographic variables, the country's budget directed towards sanitary services. On the other hand, they are directly related with the number of medical professionals and tests. The results are outstanding since these sorts of analysis haven't been taken into account to adjust the safety measures and politics adopted by various countries in order to counter the Virus' spread.

For Colombia specifically, the Ministry of health, in conjunction with Jhon Hopkins University, carried out a validation of the best indicators to trace Covid-19's behavior, which feature [12]:

- *Number of effective reproduction (Rt)*: People susceptible of being infected by a person that possess the virus.
- *Morbidity (M)*: Evolution of new cases and their accumulative tendency.
- *Lethality (L)*: Percentage of people that has died due to COVID-19.
- *General Mortality (MG)*: Considers the evolution of mortality by any cause.
- *Duplication days (d)*: Days it takes for the cases to double.
- *Positivity*: Percentage of positive samples regarding the total processed samples.
- *UCI Hospitalization*: Percentage of intensive care beds used regarding the total existing beds, not taking into account whether the bed is used by a patient with Covid-19.
- *Mobility*: Percentage of capacity used in massive transport systems throughout the country's cities.
- *Physical transactions*: Percentage of presential transactions taking place throughout the financial system in ATMs, offices, and commercial establishments.

In the previous indicators, a factor tied to population and populational density wasn't taken into consideration, as a purpose for this study these variables, and their association with the effects of the pandemic, are considered.

On the other hand, it establishes the necessity to group together information that allows the development of corresponding analysis. Precedents of Clustering can be found in Villavicencio [13], where it's applied, not to find an algorithm, instead to split logically the data in groups that could be processed.

II. Metodology

The study type is explorative since it searches to answer a hypothesis regarding the behavior of data and phenomena. In accordance with the information type, the investigation identifies itself as quantitively, although certain qualitative variables were used to develop further analysis of the results.

As for the data collection and organization, it was approached through four phases: data recompilation, selecting variables for analysis and constructing regression models to determine the correlation between data.

A. *Data collection*

For the first phase the Systematic Literature Review (SLR) was used, making use of Google® search engines to obtain data for secondary sources regarding the pandemic's behavior. To guarantee homogeneous data of various countries, data was compiled during the period between 10-03-2020 and 15-08-2020, making use of the search equation CODIV-19 AND America AND deaths AND cases AND CIFRAS OR stadistics. The recompiled information was double checked in order to delete any unreliable or repeated sources, guaranteeing the validity of the sources for the analyzed territories.

B. *Clustering*

In the second phase the data went through a clustering process by region classifying them under North American cluster, Central American cluster, and South American cluster. To accomplish this a relationship was established between Infected and three independent variables, those being Number of inhabitants, PIB per Capita and the flights per every country. With that information the coefficients of correlation and equations were determined, with this construct it was determined the levels of infection that could be reached by the different regions, with an average percentage to determine the number of deaths.

C. *Variable selection and correlation*

For the third phase, variable selection and correlation, the Excel® tool was used in order to organize the data and the correlation of the different variables, such as the levels of infection, PIB per capita, number of inhabitants per region, thousands of flights using multiple regression. Based on the results an equation for the infection forecast was found, taking into account correlation indicators between different variables. The methods used to find the correlations between the selected variables were lineal regression, multiple regression, specifying dependent and independent variables in both cases.

D. *Construction of Regression models*

In the fourth and final phase, the regression models were developed based on the correlation of the data, and through that an infection forecast was determined in every region.

III. Results

For the study of the correlation of scalar demographic variables the following were taken into account:

- Number of inhabitants per country.
- Populational density.
- ICRS Health service quality indicator.
- PIB per Capita.
- International flights (per thousands).

In terms of Covid indicators, the following were considered:

- Number of infected per country.
- Number of deceased.

Throughout development it was established to only use sovereign countries since various states associated or colonies of other countries were present, such as Puerto Rico being an associate with the United States or Curazao being dependent of Holland. Data for the development of the study was taken on the 15th of August 2020.

Table 1. Number of inhabitant's vs Infected per country.

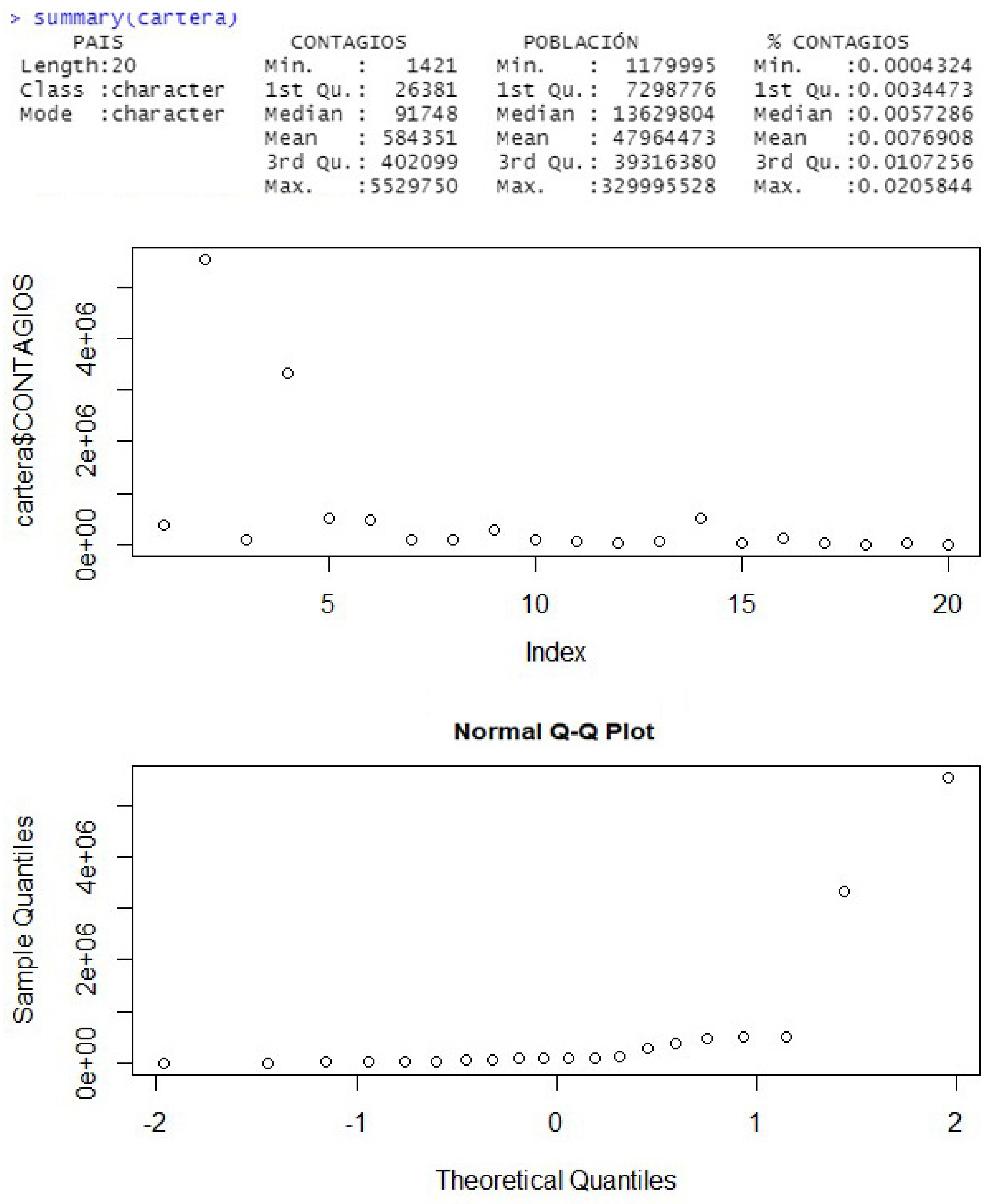
Country	Infected	Population	% Infected
Chile	383 902	18 650 114	2.1%
United States	5 529 750	329 995 528	1.7%
Panamá	79 402	5 005 246	1.6%
Brazil	3 317 832	211 823 665	1.6%
Peru	516 296	33 050 325	1.6%
Colombia	456 689	50 220 856	0.9%
Bolivia	97 950	11 969 649	0.8%
Dominican Republic	85 545	10 606 865	0.8%
Argentina	289 100	45 030 748	0.6%
Ecuador	100 688	17 080 778	0.6%
Honduras	49 467	8 893 259	0.6%
Costa Rica	27 737	6 172 543	0.4%
Guatemala	62 313	15 289 958	0.4%
Mexico	511 369	128 166 749	0.4%
El Salvador	22 314	6 356 670	0.4%
Canada	121 889	37 411 590	0.3%
Paraguay	22 314	7 612 812	0.3%
Cuba	3 229	1 179 995	0.3%
Haiti	7 810	11 485 800	0.1%
Uruguay	1 421	3 286 314	0.0%

Source: [15].

For development we started with the correlation between infected per country and the population (Table 1), the population was identified as the independent variable while the number of infected per country was identified as the dependent variable, linear regression was used for the correlation.

From the previous table we can observe that the countries with higher percentage of infection to date are Chile, United States, Panama, Brazil, Peru, which, especially in the case for US and Brazil, belonging to countries with a high number of inhabitants. Taking into account that the variables being used are scalar, a simple regression is a tool tat can offer information about the data's behavior.

Previously an analysis regarding the data was made, where we can observe the following behavior for the every one of the following variables (Fig. 1):



```
> summary(cartera)
     PAIS              CONTAGIOS          POBLACIÓN            % CONTAGIOS     
 Length:20          Min.    :   1421   Min.    :  1179995   Min.    :0.0004324  
 Class :character   1st Qu.:  26381   1st Qu.:  7298776   1st Qu.:0.0034473  
 Mode  :character   Median :  91748   Median : 13629804   Median :0.0057286  
                    Mean   : 584351   Mean   : 47964473   Mean   :0.0076908  
                    3rd Qu.: 402099   3rd Qu.: 39316380   3rd Qu.:0.0107256  
                    Max.   :5529750   Max.   :329995528   Max.   :0.0205844  
```

Fig. 1. Data behavior involving infected.
Source: Authors.

Considering the lack of collinearity in regards to the variables, the results of the regression are as follows (Table 2):

Table 2. Infected Lineal Regression results vs Lineal population.

Correlation Coefficient	0.963633972
Determination Coefficient R^2	0.928590432
R^2 adjusted	0.927162241
Common error	238896.8553
Observations	52

Source: Authors.

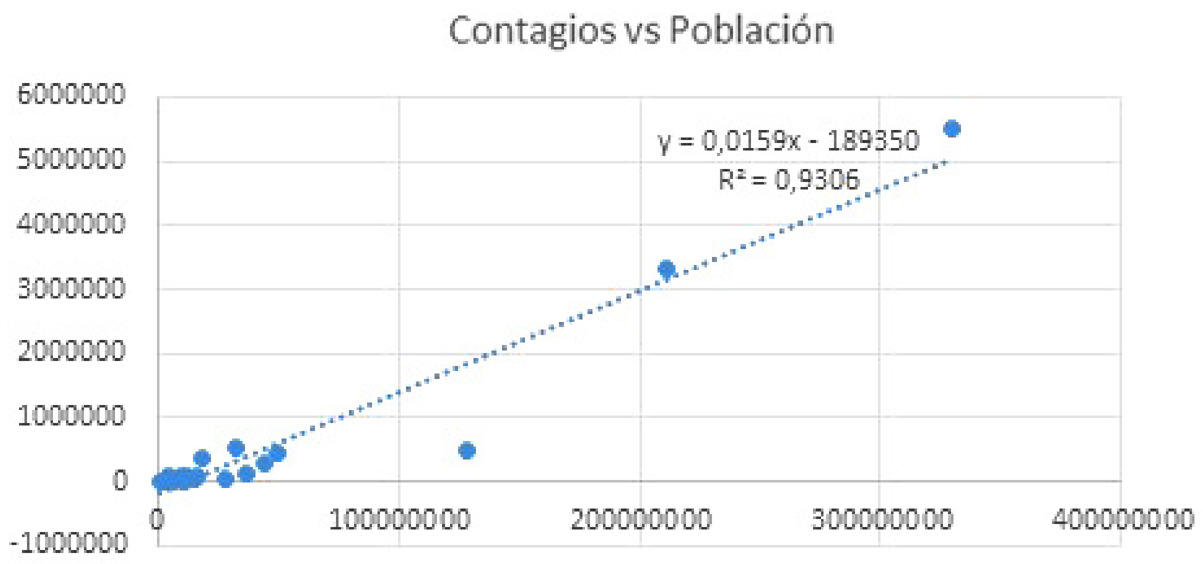


Fig. 2. Lineal Regression Infected vs Population.
Source: Authors.

Column F, from Table 3, showcases the statistic value of testing, and the probability column showcases the lowest level of significance to reject the null hypothesis of equal means. If probability is lower than 0.05 it rejects the null hypothesis, in this case it's higher, thus we accept that the infected are associated to the volume of the population (number of inhabitants).

Table 3. Infected variance analysis vs Population.

Variance Analysis			
	Degree of freedom	*F*	Critical value of *F*
Regression	1	650,1862	2,56828E-30
Remains	50		
Total	51		
	Coefficients	Probability	Below 95%
Interception	–189350.2	0.053446622	–139902.757
Variable X 1	0.01522	2.56828E-30	0.014025368

Source: Authors.

The equation for the regression (1) is represented in the following manner:

$$y = -189350{,}236 + 0{,}01522\,x \tag{1}$$

Within the results obtained its possible to observe a strong correlation of the data due to the correlation of 0.964, however the critical value of *F* is rather small which validates the use of the equation as a forecast model. Even though the relation is high there can exist other variables that can influence this result, thus the relation between populational density and the level of infection is presented, with the independent variable being the population density and the infection.

Then, the connection between infections originating from Covid 19 and the populational density, the data is organized upward based on the cases of infection (Table 4).

Table 4. Populational density vs Infections.

Countries	Covid	Density
United States	5 529 750	34.5
Brazil	3 317 832	24.1
Peru	516 296	23.9
Mexico	511 369	62.7
Colombia	456 689	41.3

Countries	Covid	Density
Chile	383 902	23.3
Argentina	289 100	15.7
Canada	121 889	3.5
Ecuador	100 688	56.7
Bolivia	97 950	9.9
Dominican Republic	85 545	217.9
Panama	79 402	49.1
Guatemala	62 313	139.4
Honduras	49 467	79
Venezuela	31 381	33.7
Costa Rica	27 737	95.3
El Salvador	22 314	292.6
Puerto Rico	10 730	389.76
Paraguay	9 022	16.8
Haiti	7 810	377.8
Nicaragua	4 115	45.7
Cuba	3 229	100.8
Surinam	2 838	3.57
Uruguay	1 421	19
Trinidad and Tobago	426	238

Source: [15].

In Table 5, the values obtained from the variance are showcased, note how the correlation coefficient Is below 0.5, which indicates that there is no correlation between populational density and infections caused by Covid 19, a possible explanation might involve mass prevention of the virus would result in avoidance of its infection.

Table 5. Density regression results vs Covid 19 infections.

Regression statistics	
Multiple Correlation Coefficient	0.173537668
Determination Coefficient R^2	0.030115322
R^2 adjusted	0.009909391
Common Error	897.391
Observations	50

Source: Authors.

The correlation's calculation has a result of 0.2 which implies a low value of correlation between the populational density and the level of infection, thus it's not viable to develop a regression equation (Table 6).

Table 6. Density variance analysis vs Covid 19 Infection.

Variance analysis		
	Degree of freedom	Critical value of F
Regression	1	0.228114531
Remains	48	
Total	49	
	Coefficients	Below 95%
Interception	357883.9353	31958.53308
Variable X 1	–597.577067	–1581.75309

Source: Authors.

Following, the contents of Table 7 establish a connection between ICRS and the amount of infections per country, the ICRS defined as the compound index of health results. The Bloomberg index of OMS is present

Table 7. Infection statistics vs ICRS.

Countries	Infection	Icrs
Canada	121 889	100
United States	5 529 750	92
Uruguay	1 421	81
Costa Rica	27 737	79
Chile	383 902	76
Cuba	3 229	76
Argentina	289 100	72
México	511 369	63
Panama	79 402	62
Paraguay	9 022	57
Brazil	3 317 832	56
Colombia	456 689	56
Ecuador	100 688	56
Dominican Republican	85 545	50
Honduras	49 467	50
El Salvador	22 314	50
Guatemala	62 313	44
Peru	516 296	42
Bolivia	97 950	24
Haiti	7 810	0

Source: [16].

Table 8. Infection regression results vs ICRS.

Regression Statistics	
Multiple Correlation Coefficient	0.286191896
Determination Coefficient R^2	0.081905801
R^2 adjusted	0.030900568
Common Error	1351213.604
Observations	20

Source: Authors.

Table 9. Infection variance analysis vs ICRS.

Variance Analysis			
	Degree of freedom	*F*	Critical value of *F*
Regression	1	1.605831324	0.221229133
Remains	18		
Total	19		
	Coefficients	Probability	Below 95%
Interception	–435872.4487	0.618195746	–2241464.467
Variable X 1	17193.23269	0.221229133	–11311.53562

Source: Authors.

Within the analysis we can observe that coefficient R adjusted is 0.081905 which implies that the connection between infection variables and the health quality index is null. Statistic wise countries such as the US, Canada and Brazil which possess a high index of health are some of the countries with the biggest amount of cases of infection, meanwhile countries with lower indices also present lower infection rates. It's important to note that the number of tests provided by the country might distort the statistics.

Then, it proceeds to form a correlation of the variables infection vs people transported by aerial transports, showcased in the following Table 10 are the results of the correlation:

Table 10. Infection statistics vs Flights (thousands) per country.

Countries	Infected	Flights (Thousands)
United States American	5 529 750	889 022
Canada	121 889	121 889
Brazil	3 317 832	102 109
Mexico	511 369	64 529
Colombia	456 689	33 704
Chile	383 902	19 519
Argentina	289 100	18 081
Peru	516 296	17 758
Panama	79 402	12 939
Ecuador	100 688	5 365
Bolivia	97 950	4 122
El Salvador	22 314	2 545
Costa Rica	27 737	1 948
Uruguay	1 421	563
Cuba	3 229	561
Paraguay	22 314	560
Honduras	49 467	411
Guatemala	62 313	146
Dominican Republic	85 545	111
Haiti	7 810	12

Source: [15].

Table 11. Infection regression results vs Flights by Thousand.

Regression Statistics	
Multiple Correlation Coefficient	0.892087715
Determination Coefficient R^2	0.795820491
R^2 adjusted	0.784477185
Common Error	637081.2013
Observations	20

Source: Authors.

Table 12. Infection variance analysis vs Flights by thousand.

Variance Analysis		
	Degree of freedom	Critical value of *F*
Regression	1	1.27E-07
Remains	18	
Total	19	
	Coefficients	Below 95%
Interception	181932.3743	–133918.4567
Variable X 1	6.210669634	4.652873464

Source: Authors.

It is noted a correlation coefficient of 0.89, which indicates a strong connection between air transport and infection, we can conclude the following; aerial transport of passenger might be a variable for high levels of infections, which justifies one of the objectives for various governments was to cancel national and international flights, mainly to avoid further spread of the virus. We can define it with (2):

$$y = 181932.3743 + 6.210669634x1 \qquad (2)$$

A. *Multiple regression analysis*

for the development of the multiple regression analysis it was decided to combine the PIB per capita from 2019 (economic type variable) with the number of inhabitants and the level of infection, with the purpose to establish if there exists any connection from an economic standpoint. Below is the table with the results of the analysis.

Table 13. Infection statistics vs PIB per capita and number of inhabitants.

Countries	Infected	Pib Per Capita	Population
USA	5 529 750	65 118	329 995 528
Brazil	3 317 832	8 717	211 823 665
Peru	516 296	6 977	33 050 325
México	511 369	9 863	128 166 749
Colombia	456 689	6 432	50 220 856
Chile	383 902	14 896	18 650 114
Argentina	289 100	10 006	45 030 748
Canada	121 889	46 194	37 411 590
Ecuador	100 688	6 183	17 080 778
Bolivia	97 950	3 552	11 969 649
Republican Dominican	85 545	8 282	10 606 865
Panamá	79 402	15 731	5 005 246
Guatemala	62 313	4 620	15 289 958
Honduras	49 467	2 574	8 893 259
Costa Rica	27 737	12 238	6 172 543
El Salvador	22 314	4 187	6 356 670
Paraguay	22 314	5 415	7 612 812
Haiti	7 810	755	11 485 800
Cuba	3 229	8 821	1 179 995
Uruguay	1 421	16 190	3 286 314

Source: [15].

Table 14. Multiple regression Infected vs population and PIB per capita .

Regression Statistics	
Multiple Correlation Coefficient	0.96666375
Determination Coefficient R^2	0.9344388
R^2 adjusted	0.92672572
Common Error	371470.177
Observations	20

Source: Authors.

Table 15. Infection variance analysis vs population and PIB per capita .

Variance Analysis		
	Degree of freedom	Critical value of F
Regression	2	8.74E-11
Remains	17	
Total	19	
	Coefficients	Below 95%
Interception	–219248.241	–449109.8121
Variable X 1	6.1405665	–8.980253
Variable X2	0.01511054	0.012297051

Source: Authors.

According to the data we can observe that the correlation coefficient is 0.965 which implies a strong correlation of the studied variables. There is a higher number of cases in countries that present a high economic level, such being the case of US and Brazil. The mathematical model for the multiple regression is the following:

$$Y = -219248.241 + 6.1405655X1 + 0.0151\ X2 \qquad (3)$$

Y being = Dependent variable number of cases

X1 = PIB Per Capita value per country

X2 = Number of inhabitants per country.

Next, the multiple regression possesses a dependent variable (Infections) and three independent variables, those being PIB per Capita, Population and number of flights (by thousand).

Table 16. Infection Statistics vs PIB per Capita, Number of inhabitants and Flights by thousand.

Countries	Infection	Pib Per capita	Population	Flights (thousand)
USA	5 529 750	65 118	329 995 528	5 529 750
Mexico	511 369	9 863	128 166 749	64 529
Canada	121 889	46 194	37 411 590	121 889
Dominican Republic	85 545	8 282	10 606 865	111
Panama	79 402	15 731	5 005 246	12 939
Honduras	49 467	2 574	8 893 259	411
Guatemala	62 313	4 620	15 289 958	146
El Salvador	22 314	4 187	6 356 670	2 545
Haiti	7 810	755	11 485 800	12
Cuba	3 229	8 821	1 179 995	561
Costa Rica	27 737	12 238	6 172 543	1 948
Brazil	3 317 832	8 717	211 823 665	102 109
Peru	516 296	6 977	33 050 325	17 758
Colombia	456 689	6 432	50 220 856	33 704
Chile	383 902	14 896	18 650 114	19 519
Argentina	289 100	10 006	45 030 748	18 081
Ecuador	100 688	6 183	17 080 778	5 365
Bolivia	97 950	3 552	11 969 649	4 122
Paraguay	22 314	5 415	7 612 812	560
Uruguay	1 421	16 190	3 286 314	563

Source: Authors.

The results from the regression are as follows (Table 17):

Table 17. Infection multiple regression results vs PIB per Capita, Population and Flights by thousand.

Regression Statistics	
Multiple Correlation Coefficient	0.97506441
Determination Coefficient R^2	0.9507506
R^2 adjusted	0.94151633
Common Error	331868.067
Observations	20

Source: Authors.

Table 18. Multiple regression variance analysis vs population, PIB per capita and flights by thousand.

Variance Analysis		
	Degree of freedom	Critical value of F
Regression	3	1.13E-10
Remains	16	
Total	19	
	Coefficients	Below 95%
Interception	–51321.308	–309176.9448
Variable X 1	–5.6637407	–23.05349225
Variable X 2	0.01286134	0.009595049
Variable X 3	0.30823405	0.024385496

Source: Authors.

The multiple correlation coefficient is 0.97506441 which indicates a direct connection between the variables and establishes the following linear equation (4):

$$Y = -51321.3079 - 5.667X1 + 0.01286X2 + 0.308X3 \quad (4)$$

B. *Clustering process*

A grouping technique was utilized to explore the data attempting to maximize the similarities between the elements of the class and minimize the similarities between groups. Throughout the clustering process the countries were classified by region: North American, Central American (plus the Caribbean islands) and South American.

1) *North America Cluster*

Table 19. Infection statistics Vs PIB Per capita and Cluster North America population.

Countries	Infected	Pib Percapita	Población	Flights in Thousands
USA	5 529 750	65 118	329 995 528	5 529 750
Mexico	511 369	9 863	128 166 749	64 529
Canada	121 889	46 194	37 411 590	121 889

Source: Authors.

For this cluster the Infection (y), PIB per capita (X1) and Population (X2) were used, and yielded the following results (Table 20):

Table 20. Infection regression results vs PIB Per capita and Cluster North America population.

Regression Statistics	
Multiple Correlation Coefficient	1
Determination Coefficient R^2	1
R^2 adjusted	65535
Common Error	0
Observations	3

Source: Authors.

From the analysis of variance, we obtained a strong correlation coefficient, above 0.5. With R2, we can determine that the model adjusts itself to the data (Table 21).

Table 21. Infection variance analysis Vs PIB Per capita and Cluster North America population.

Variance Analysis		
	Degree of freedom	Critical value of F
Regression	2	0
Remains	0	
Total	2	
	Coefficients	Below 95%
Interception	–1905568.2	–1905568.2
Variable X 1	30.5196402	30.5196402
Variable X 2	0.01650913	0.01650913

Source: Authors.

From the results we can gather a correlation of 1, which indicates a direct connection between the variables, resulting in the following equations (5):

$$Y = -1905568.197 + 30.51964018X1 + 0.016X2 \qquad (5)$$

In accordance with the multiple regression equation North America's cluster reached an infection level of 9 721 831 inhabitants.

2) *Central American (& Caribbean Islands) Cluster*

The following table selects countries of Central America (& the Caribbean Islands), with the purpose to analyze the infection and establish if there is any correlation with nearby territories (Knn) (Table 22).

Table 22. Infection statistics vs PIB per Capita, Population and flights by thousand in Central America.

Countries	Infected	Pib Percapita	Population	Flights in Thousands
Dominican Republic	85 545	8282	10606865	111
Panama	79 402	15731	5005246	12939
Honduras	49 467	2574	8893259	411
Guatemala	62 313	4620	15289958	146
El Salvador	22 314	4187	6356670	2545
Haiti	7 810	755	11485800	12
Cuba	3 229	8821	1179995	561
Costa Rica	27 737	12238	6172543	1948

Source: Authors.

For the Central American cluster there was one independent variable (infection) and three dependent variables, those being X1 (PIB per Capita), X2 (population) and X3 (Flights per year by thousand unit). The results from the lineal regression are as follows:

Table 23. Infection regression results vs PIB per Capita, Population, and Flights by thousand unit in Central America.

Regression Statistics	
Multiple Correlation Coefficient	0.7956783
Determination Coefficient R^2	0.6331039
R^2 adjusted	0.3579319
Common Error	25386.36
Observations	8

Source: Authors.

In accordance with the Value obtained from F, the variance of the factors doesn't appear to present any significant effect in the overall result (Table 24).

Table 24. Infection variance analysis vs PIB per Capita, Population and flights by thousand in Central America.

Variance Analysis		
	Degree of freedom	Critical value of F
Regression	3	0.219018715
Remains	4	
Total	7	
	Coefficients	Below 95%
Interception	–36702,63	–132478.9296
Variable X 1	3,8556841	–4.4163367
Variable X 2	0,0057832	–0.0014193
Variable X 3	1,8749147	–6.866583575

Source: Authors.

Within the correlation coefficient we can observe a value of 0.79 which indicates a direct connection of medium value. The multiple regression equation is defined as (6):

$$Y = -36702.6297 + 3.85568X1 + 0.00578319X2 + 1.87491474X3 \tag{6}$$

According to the forecast resulting from the equation, an infected level of 668140 for the Central American & Caribbean region is estimated.

3) *South American Cluster*

Table 25. Infection statistics vs PIB per Capita, Population and flights by thousand in South America Cluster.

Countries	Infected	Pib Percapita	Population	Flights Thousands
Brazil	3 317 832	8 717	211 823 665	102 109
Peru	516 296	6 977	33 050 325	17 758
Colombia	456 689	6 432	50 220 856	33 704
Chile	383 902	14 896	18 650 114	19 519
Argentina	289 100	10 006	45 030 748	18 081
Ecuador	100 688	6 183	17 080 778	5 365
Bolivia	97 950	3 552	11 969 649	4. 122
Paraguay	22 314	5 415	76 12 812	560
Uruguay	1 421	16 190	3 286 314	563

Source: Authors.

On the following table are the results from the lineal regression (Table 26).

Table 26. Infection regression results vs PIB per Capita, Population and flights by thousand in South America Cluster.

Regression Statistics	
Multiple Correlation Coefficient	0.989070992
Determination Coefficient R^2	0.978261427
R^2 adjusted	0.965218284
Common Error	194998.2521
Observations	9

Source: Authors.

Table 27. Infection variance analysis vs PIB per Capita, Population and flights by thousand in South America Cluster.

Variance Analysis		
	Degree of freedom	Critical value of F
Regression	3	0.000140834
Remains	5	
Total	8	
	Coefficients	Below 95%
Interception	–217637.9657	–643803.084
Variable X 1	9.599893018	–33.57114109
Variable X 2	0.014013166	–0.003010874
Variable X 3	3.990495374	–30.68113114

The correlation coefficient is 0.9807 which implies a strong and direct connection of the correlation between variables, and establishes the following regression equation (7):

$$Y = -217637.966 + 9.59989302X1 + 0.01401317X2 + 3.99049537X3 \qquad (7)$$

According to the established forecast based on the multiple regression equation South America will reach a level of infection estimated to be around 6 926 041 people (Fig. 3).

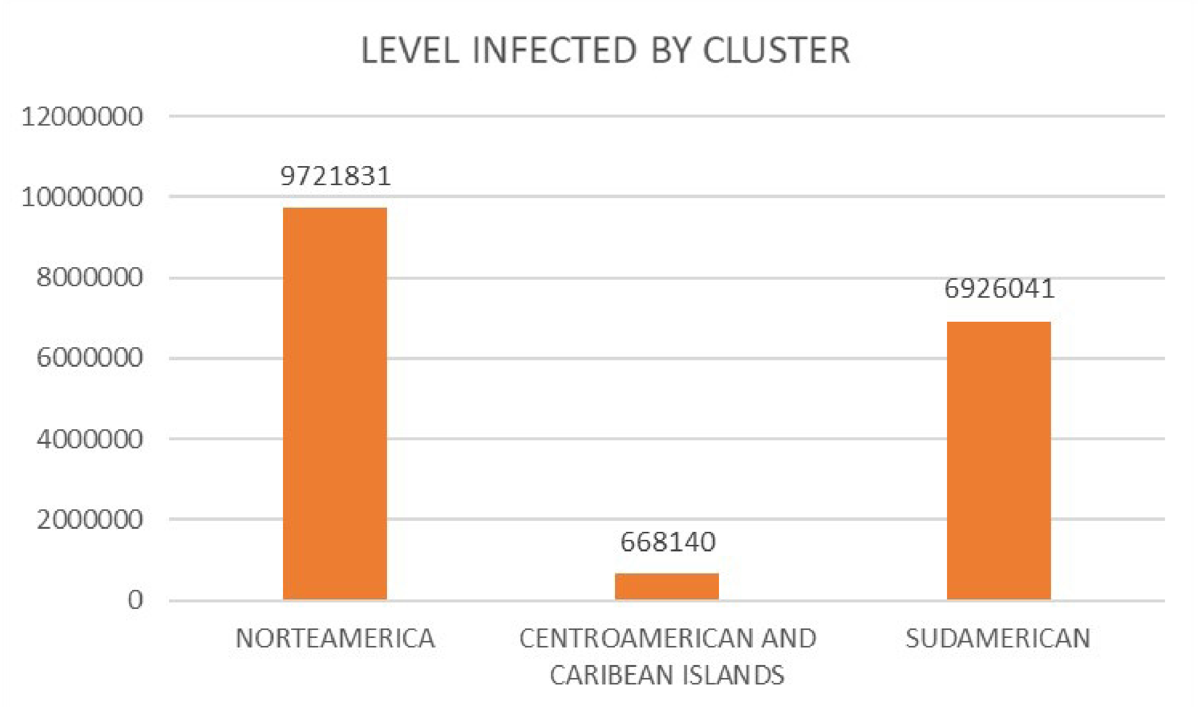


Fig. 3. Infection level per cluster
Source: Authors.

When adding up the cluster it results in an estimated infection level of 17 316 012 inhabitants for the American continent and deaths averaging 3% resulting in a value of 519 480 deaths.

According to Fig. 4 we can observe that by the Month of August (2020) the American countries represent the largest figures of COVID confirmed cases, and countries present similar figures in regards to the infection rates in concordance with the clusters previously established (North America, Central America and South America).

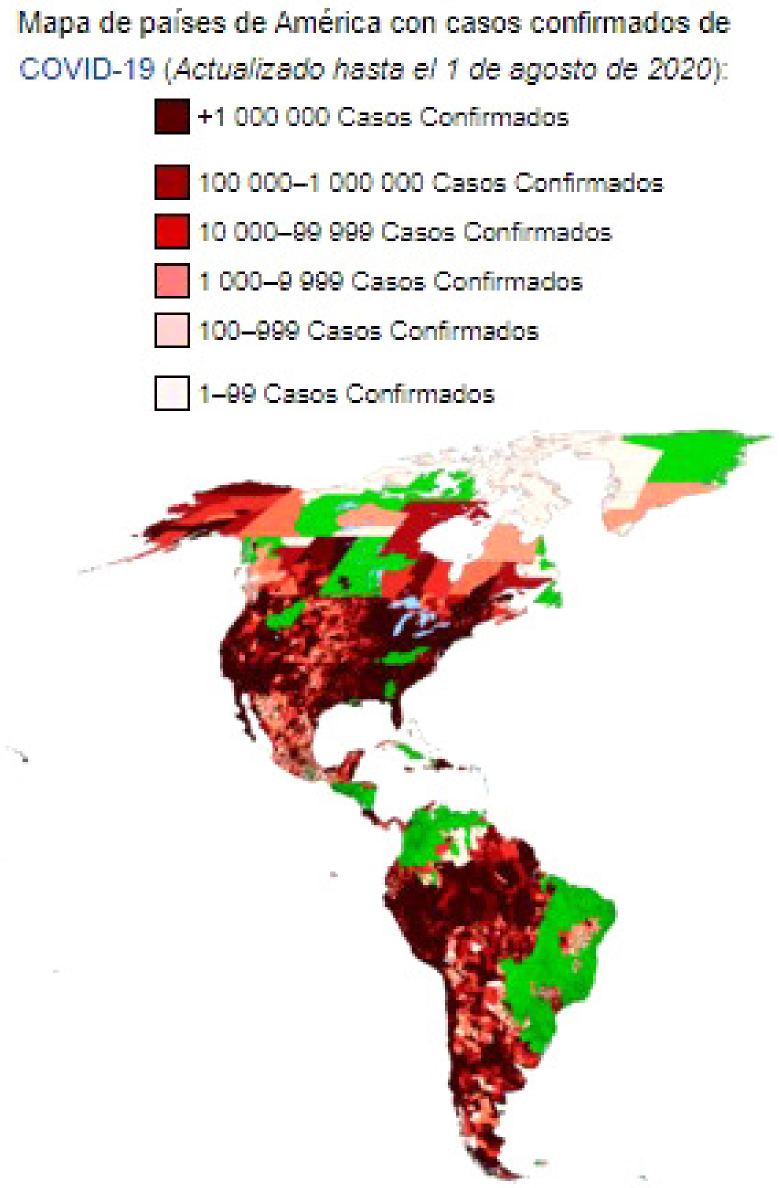

Mapa de países de América con casos confirmados de
COVID-19 (Actualizado hasta el 1 de agosto de 2020):
+1 000 000 Casos Confirmados
100 000–1 000 000 Casos Confirmados
10 000–99 999 Casos Confirmados
1 000–9 999 Casos Confirmados
100–999 Casos Confirmados
1–99 Casos Confirmados

Fig. 4. World Health Organization's situation reports. Johns Hopkins University CSSE, The Centers for Disease Control and Prevention, New York Times, CNBC.
Source:

IV. Conclusions

The findings suggest that connectivity plays a larger role in the propagation of the COVID-19 pandemic than populational density. Large metropolitan areas with a high number of spaces deeply connected via the economy, social and transport are the most vulnerable.

As stated by the study there appears to be a strong correlation of variables such as the amount of inhabitants, the amount of flights and the economic activity reflected by the PIB per capita of the continent's countries, clustering the data by regions, which implies that the measures taken in regards to the border closure resulted in an alternative for controlling the propagation at a global level.

Indicators of health service quality and population density of the countries were taken, but no correlation of the data was defined, which implies that several countries with low levels of ICRS did not perform enough tests to define the number of infected people.

The economic level and the amount of flights had a correlation regarding the level of infection. This is reflected in the first measures used by the governments; this being quarantining the area (such as prohibiting travel and certain economic activities) to minimize the infection rate.

Studies found in literature use other statistical methods to develop the correlation with the variables [14], [9] and [11], some of which stand out such as the use of R-square values, Spearman correlation coefficient an the statistic analysis ANOVA, which establishes a need to develop other correlational analysis with techniques that allow the passing of time, and thus show estimates about the variable's behavior in the pandemic's future.

Variables such as population, PIB per Capita and flights generate a high relation with the infection level, but it's necessary to point out that this doesn't imply the causality of the infection since correlational analysis don't have the reach necessary to determine that information.

Financing

This investigation belongs to the research project known as "*Diseño de una metodología para la gestión de flujo de pacientes en servicios de salud basado en estrategias lean y patrones de proceso de negocios",* Answers to the process betterment line belonging to the Organizational processes investigation group (GIPO) in business engineering approved by the investigation committee from the Industrial Engineering faculty of the San Tomas University.

References

[1] Real Académica Española, *Diccionario de la lengua española.* MD, ES: RAE, Oct. 2019. Disponible en https://dle.rae.es/

[2] T. N. Jilani, R. T. Jamil & A. H. Siddiqui, "H1N1 Influenza," in, *StatPearls [Internet].* Treasure Island, FL: StatPearls Publishing, 2020.

[3] S. Hamidi, S. Sabouri & R. Ewing, "Does Density Aggravate the COVID-19 Pandemic?," *J Am Plan Assoc*, vol. 86, no. 4, pp. 495–509, 2020. https://doi.org/10.1080/01944363.2020.1777891

[4] B. Ather, T. M. Mirza & P. F. Edemekong, "Airborne Precautions," in, *StatPearls [Internet].* Treasure Island, FL: StatPearls Publishing, 2020.

[5] P. S. Peixoto, D. Marcondes, C. Peixoto & S. M. Oliva, "Modeling future spread of infections via mobile geolocation data and population dynamics. An application to COVID-19 in Brazil," *PLoS One*, vol. 15, no. 7, Jul. 2020. https://doi.org/10.1371/journal.pone.0235732

[6] D. Rosselli, "Epidemiología de las pandemias," *Rev Medicina*, vol. 42, no. 2, pp. 168–174, Jul. 2020. Disponible en https://revistamedicina.net/ojsanm/index.php/Medicina/article/view/1511

[7] OPS, "Por qué es importante el desglose de datos durante una pandemia," *paho.org*, 2020. Recuperado de https://www.paho.org/ish/images/docs/Data-Disaggregation-Factsheet-Spanish.pdf

[8] OPS, "*Pandemia de COVID-19: estadísticas sobre el acceso a la BVS y el alcance de la cooperación técnica de BIREME*, *Boletín Bireme*, no. 42, 2020. Disponible en https://boletin.bireme.org/2020/04/01/pandemia-de-covid-19-estadisticas-sobre-el-acceso-a-la-bvs-y-el-alcance-de-la-cooperacion-tecnica-de-bireme/

[9] P. G. Ruiz Mamani, W. C. Morales-García, M. White & M. S. Marquez-Ruiz, "Properties of a scale of concern for COVID-19: Exploratory analysis in a Peruvian sample," *Med Clin*, vol. 255, no. 12, pp. 535–537, Dec. 2020. https://doi.org/10.1016/j.medcli.2020.06.022

[10] S. Hamidi, R. Ewing & S. Sabouri, "Longitudinal analyses of the relationship between development density and the COVID-19 morbidity and mortality rates: Early evidence from 1,165 metropolitan counties in the United States," *Heal Place*, vol. 64, no. 2, pp. 102378–102378, Jul. 2020. https://doi.org/10.1016/j.healthplace.2020.102378

[11] A. Medeiros de Figueiredo, A. Daponte, D. C. Moreira Marculino de Figueiredo, E. Gil-García & A. Kalache, "Case fatality rate of COVID-19: absence of epidemiological pattern," *Gac. Sanit*, vol. 35, no. 4, pp. 10–12, 2020. https://doi.org/10.1016/j.gaceta.2020.04.001

[12] República de Colombia. MinSalud, "Análisis de la epidemia de covid-19 en el país," *Boletín de Prensa No 223 de 2020*, 2020. Disponible en https://www.minsalud.gov.co/Paginas/Analisis-de-la-epidemia-de-covid-19-en-el-pais.aspx

[13] F. Velásquez & G. D. Sosa, "Aplicación de Técnicas de Clustering en Sonidos Adventicios para Mejorar la Interpretabilidad y Detección de Estertores," *INE CUC*, vol. 11, no. 1, pp. 53–62, 2015. Disponible en http://revistascientificas.cuc.edu.co/index.php/ingecuc/article/download/366/2015105

[14] I. F. Meza, A. E. Herrera & L. G. Obregón, "Determinación experimental de nuevas correlaciones estadísticas para el cálculo del coeficiente de transferencia de calor por convección para placa plana, cilindros y bancos de tubos," *INGECUC*, vol. 13, no. 2, pp. 9–17, 2017. https://doi.org/10.17981/ingecuc.13.2.2017.01

[15] GBM, Banco Mundial de la Salud, *datos.bancomundial.org*, 2019. Disponible en https://datos.bancomundial.org/indicator/NY.GDP.PCAP.CD

[16] ACCH, Asociación Colombiana de Clínicas y hospitales, *achc.org*, 2019. Disponible en http://achc.org.co

Elisa del Carmen Navarro-Romero. Industrial engineering from the Universidad del Norte (Colombia). Magister in industrial engineering with emphasis on Organizational management, from the Universidad Distrital Francisco de Paula Santander (Colombia); teacher in investigational formation. 12 years of experience with business from the real sector, with positions associated with betterment processes, budget, guaranteeing quality and logistics, in Financial companies such as the Grupo Aval Holding; Service companies such as CUMANDES-Cummins de los Andes, and production companies such as Oleoflores SA, and other of the sort. https://orcid.org/0000-0002-1825-0097

Óscar Mauricio Gelves-Alarcón. Full time teacher at the faculty of Industrial Engineering at the Universidad Militar Nueva Granada (Colombia), dedicated to educational spaces involving modern manufacture and economical engineering, with schooling in Industrial Engineering at the District University, specialized in production engineering. Master's degree in industrial directed engineering at the Universidad de Buenos Aires (Argentina) and with experience in education in various private universities in the fields of Production, investigation and logistics, with experience in the real sector as coordinator of planning at Brightstar and coordinator of logistics at Eco de los Andes. https://orcid.org/0000-0003-0557-775X

Natalia García-Corrales. Industrial engineering at the Universidad del Norte (Colombia), Magister in administration and specialist in project management at the Universidad Pontificia Bolivariana (Colombia), and specialist in health & security management at the ECCI university. Currently, professor at the faculty of industrial engineering at the Pontificia Universidad Bolivariana (Colombia), and leader of strategic management of the Environment & Technology Foundation. 12 years of experience with management system articulation with management and strategic planning, particularly with service companies. Organizational advice in organizational transformation processes. https://orcid.org/0000-0002-6866-2401