Spatial distribution of excess deaths due to COVID-19 in Ecuador: Socioeconomic disparities in Ecuadorian provinces

[redacted]

June 22, 2023

Abstract

The COVID-19 pandemic had a devastating impact on resource-limited countries like Ecuador. To understand the factors contributing to this problem, we analyzed the Excess Death Factor (EDF) in Ecuadorian provinces during the pandemic. Our study employed spatial exploration, multicollinearity assessment, and Linear Regression (LM), Spatial Autoregressive (SAR), and Conditional Autoregressive (CAR) models. We found that the unemployment rate, and the rates of White and Black populations, were significant predictors of EDF in Ecuadorian provinces during COVID-19. These findings raise important questions about the influence of these factors on the severity of the pandemic. By uncovering these relationships, our study enhances our understanding of the complex dynamics between socioeconomic factors and COVID-19 outcomes. This knowledge can inform targeted interventions and policies to mitigate the impact of future pandemics in Ecuador and similar contexts.

Introduction

The COVID-19 pandemic in Ecuador is part of the worldwide pandemic of coronavirus disease 2019 (COVID-19). The first reported case of the virus in Ecuador was confirmed on February 29, 2020, when a woman in her 70s tested positive. In April 2020, the COVID-19 pandemic in Ecuador was described as emerging as a possible "epicenter" of the pandemic in Latin America. The city of Guayaquil, in particular, became overwhelmed to the extent that bodies were left in the streets[1].

Between January 1 and September 23, 2020, Ecuador experienced a significant number of excess deaths, with a total of 36,402 excess deaths (95% CI: 35,762–36,827) or 208 per 105 population. Excess deaths refer to the number of deaths that exceed the expected or normal levels within a specific time period. It indicates an increase in mortality beyond what would typically be anticipated based on historical data or statistical projections. These deaths accounted for 171% of the expected deaths in a typical year.[2]

It is important to note that excess deaths in 2020 have been formally estimated previously for many high-income countries with high COVID-19 cases. The numbers of all-cause deaths range between 120% and 131% of expected deaths in these countries, i.e. between 20% and 31% of excess deaths, highlighting the heavy burden of COVID-19 spread directly on mortality, despite differences in demographics, social mixing patterns and health care systems. In these countries, a major fraction of excess deaths (between 67% to 80%) are attributable to COVID-19 deaths,[3-6] suggesting that the majority of excess deaths are caused directly by COVID-19 infections. In contrast, Ecuador reported a relatively small number of COVID-19 deaths13; however, a surprisingly large number of excess deaths was reported in the Our World in Data online database.[7]

This study builds upon the previous work conducted on excess deaths and the impact of COVID-19 in Ecuador. Previous research, such as the study "Excess deaths reveal the true spatial, temporal, and demographic impact of

COVID-19 on mortality in Ecuador" by Leticia Cuéllar et al, revealed the significant surge in per capita deaths in Ecuador during the early stages of the pandemic. It highlighted the existence of a large number of excess deaths, with only a fraction attributed to confirmed COVID-19 cases.[2]

However, this current study takes a different approach by focusing on identifying specific socioeconomic factors that may have contributed to the higher spread of COVID-19 in certain provinces. It aims to uncover the underlying reasons why the disease affected some regions more severely than others. While previous studies highlighted the presence of excess deaths likely linked to COVID-19, they did not thoroughly investigate the socioeconomic variables associated with the disproportionate impact.

In contrast, this study explores the relationship between excess deaths due to COVID-19 and various socioeconomic factors at the province level. By employing spatial models and analyzing province-level data, it seeks to address two key research questions: (1) whether there is spatial autocorrelation in the distribution of the Excess Death Factor (EDF) across Ecuadorian provinces, and (2) which specific socioeconomic factors, if any, are correlated with EDF in each province. This research aims to uncover new insights into the factors contributing to the higher transmission rates and severity of the disease in certain provinces, providing valuable knowledge for future mitigation strategies and interventions.

By focusing on socioeconomic reasons behind the varying rates of COVID-19 spread across provinces, this study expands the understanding of the pandemic's impact in Ecuador. Its unique approach contributes to the existing body of knowledge by delving deeper into the specific factors that may have influenced the disease's severity at a regional level.

Data

I began by collecting several comprehensive datasets relating to demographic factors across Ecuadorian provinces and the effect of COVID-19 on deaths in Ecuadorian provinces. These data include COVID death data collected in 2021 through the ecuacovid git repository[8], the distribution of ages and population across Ecuadorian provinces in 2021 from the Statistica website[9], excess death data between January 1st and September 23rd, 2020, collected from the Ecuadorian National Institute of Statistics and Census and the Ecuadorian Ministry of Government[10], and extraction from IPUMS of Ecuador demographics from the 2010 census[11].

The COVID death data provided information on COVID-related deaths in Ecuadorian provinces for the year 2021[4]. The dataset obtained from the Statistica website included data on the distribution of ages and population across Ecuadorian provinces in 2021[5], enabling insights into the age composition of the population within each province.

To examine the impact of COVID-19 on mortality rates, the dataset on excess deaths between January 1 and September 23, 2020, was acquired[6]. This data, collected from the Ecuadorian National Institute of Statistics and Census and the Ecuadorian Ministry of Government, provided valuable insights into the excess deaths associated with COVID-19 in Ecuadorian provinces.

Demographic data from the 2010 census of Ecuador were obtained from IPUMS[7]. This dataset required a personal account on the IPUMS website and a data request. It included variables related to socioeconomic factors such as ethnicity, literacy level, schooling, employment, and hours worked per week. The IPUMS dataset provides a representative sample of one-tenth of the entire Ecuadorian population at the household level.

Methods

I initiated the analysis of EDF in Ecuadorian provinces by conducting spatial exploration of EDF along with other potential explanatory variables that I had collected. This exploration involved graphing choropleths of EDF, population, unemployment, rate of population aged 60 and older, and ethnicity. The purpose was to visualize the spatial distribution of these variables in the provinces. To guide the spatial analysis, I referred to the techniques outlined in "Applied Spatial Data Analysis with R" by Bivand.[12]

To ensure the validity of the analysis, I examined the presence of multicollinearity among the explanatory variables. I started by calculating the variance inflation factor (VIF) for the rate of population aged 60 and older, the rate of unemployment, the rate of non-educated individuals, and the rates of indigenous, white, mestizo, and black populations. High VIF values indicate the presence of multicollinearity, where the coefficients' variance estimates are inflated for variables that are highly correlated. To address this issue, I removed the variables that exhibited multicollinearity, namely the rate of non-educated individuals and the rate of indigenous population. Subsequently, I proceeded with the analysis.

To model the relationship between EDF and the remaining explanatory variables, I employed linear regression (LM). Additionally, I constructed a binary neighborhood matrix, which was necessary for conducting Moran's I testing and building the Spatial Autoregressive (SAR) and Conditional Autoregressive (CAR) models. Prior to further spatial analysis, it was crucial to assess the presence of spatial autocorrelation in the residuals of the LM model using Moran's I. The residuals represent the differences between the observed values of EDF and the predicted values based on the covariates. Significant spatial autocorrelation in the residuals would indicate unexplained spatial structure or clustering that the current model failed to capture. the key assumptions is that the residuals (the differences between observed and predicted values) are independent and identically distributed (i.i.d.). If spatial autocorrelation is present in the residuals, it indicates that the residuals are not independent and violates on of the key assumptions of linear regression, that the residuals are independent and identically distributed (i.i.d.). To address this violation, alternative modeling approaches that explicitly account for spatial autocorrelation are necessary.

Conversely, if no significant spatial autocorrelation was detected in the residuals, it would suggest that the model adequately captured the spatial structure, and further spatial analysis would not be necessary.[12]

In my analysis of the EDF model for Ecuadorian provinces, no significant spatial autocorrelation was found in the residuals. However, this does not completely negate the valuable benefits of conducting spatial analysis. Spatial analysis offers a comprehensive perspective on the geographical distribution of the data. Moreover, spatial analysis facilitates the visualization and communication of findings by enabling the creation of informative maps and spatial representations. In my study, spatial analysis helps identify statistical patterns present in visible choropleths of Ecuador. These visualizations enhance the understanding of the data and facilitate effective decision-making processes that may be due to the various socioeconomic factors in my study.[13,14]

Next, I fitted SAR and CAR models, accounting for spatial structure and dependencies, using the remaining explanatory variables and EDF as the response variable. The previously fitted LM model served as a baseline model that did not account for spatial dependence. SAR models take into account the spatial autocorrelation of the EDF, meaning that the EDF in one province is influenced by the EDF values in its neighboring provinces. On the other hand, CAR models not only consider spatial autocorrelation but also incorporate the spatial dependence of the error term. while both SAR and CAR models take spatial relationships into account, the SAR model focuses on the memoryless property of EDF, the response variable, whereas the CAR model considers the memoryless property for both EDF and the error terms. In the CAR model, this means that both EDF and the error term have spatial dependencies for a province that influence the values in neighboring provinces. These models assume that the error terms of neighboring provinces are correlated, providing a more comprehensive understanding of the spatial patterns and relationships associated with the EDF in Ecuadorian provinces.

The generalized formula for both SAR and CAR models are given below:

SAR Model:

$$Y = X'\beta + \lambda W(Y - X'\beta) + \varepsilon$$

where:

Y represents the dependent variable (EDF), X represents the matrix of independent variables, with X' denoting its transpose, including rate_60, UNEMP_rate, mestizo_rate, white_rate, and black_rate, β represents the vector of regression coefficients, λ is the spatial autocorrelation parameter, which is estimated from the data, W represents the neighborhood matrix or contiguity matrix, capturing the spatial relationships between units, and ε represents the error term.

CAR Model:

$$Y_i \,|\, Y_{j\sim i} \sim N(X\beta + \lambda W(Y - X\beta) + ... + \varepsilon)$$

where:

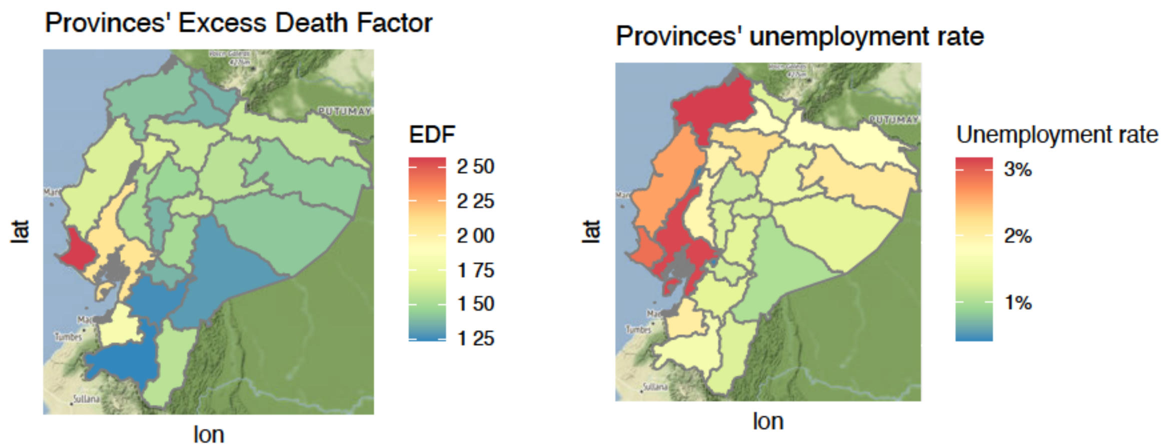
Y_i represents the dependent variable (EDF), $Y_{j\sim i}$ indicates the dependent variable in the neighbors of unit i, X is the matrix of independent variables, including rate_60, UNEMP_rate, mestizo_rate, white_rate, and black_rate, β is the vector of coefficients associated with the independent variables, λ is the spatial autoregressive coefficient that captures the spatial dependence, W is the spatial weights matrix that represents the spatial relationships between neighboring units, and ε is the error term.

In the model, $(...)$ represents any additional parameters or assumptions specific to the model that are not explicitly mentioned in this equation.

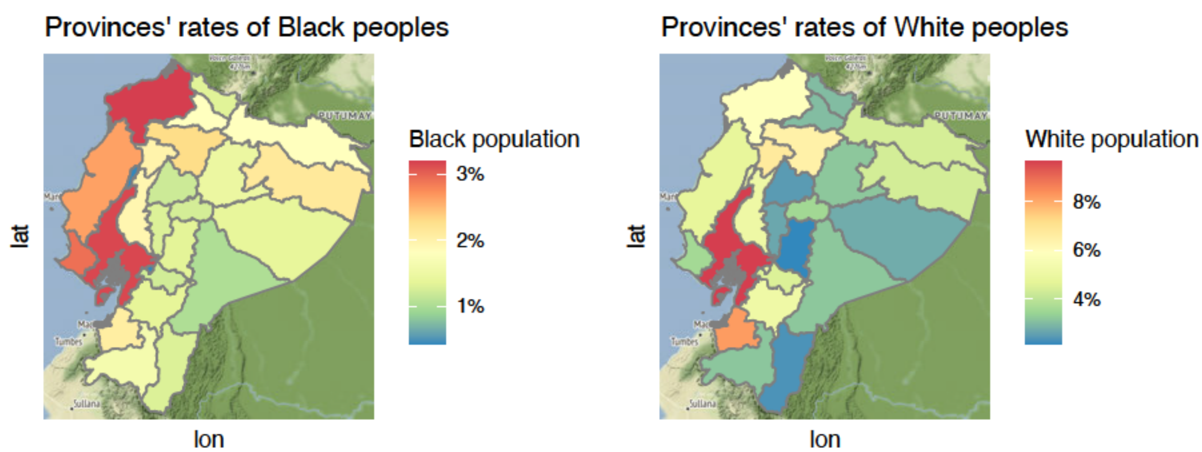
To compare the SAR and CAR models, I utilized the Akaike Information Criterion (AIC). Lower AIC values indicate better model fits, striking a balance between goodness of fit and model complexity. Minimizing the AIC value guides the selection of a model that accurately represents the data while avoiding overfitting with an excessive number of parameters. By selecting the model with the lowest AIC, I prioritize a good fit to the data while considering the model's complexity. This approach ensures a trade-off between accurately representing the data and avoiding excessive complexity. [12]

After identifying the most suitable model, I will proceed to analyze the results obtained from this model. These results will be discussed, highlighting the significance of various socioeconomic factors on EDF.

Results



Figures 1 & 2 : Choropleths of EDF and Unemployment rate



Figures 3 & 4 : Choropleths of Black and White population percentage

Initially, we performed spatial exploration by graphing choropleths of EDF and potential explanatory variables, including population, unemployment rate, rate of population 60 and older, and ethnicity. These visualizations provided a preliminary understanding of the spatial patterns and relationships in the data.(Figures 1-4) It is evident that there appears to be a larger EDF on the west coast of the country in where there appears to be a a higher rate of unemployment, black peoples, and white peoples.

Next, we began the construction of a Linear Model (LM) utilizing provinces' ages over 60 rate, unemployment rate, illiterate rate, persons uneducated rate, total population, mestizo rate, white rate, black rate, and indigenous rate as explanatory variables. These four races were used as they are the most common ethnic groups within Ecuador.[14]

To address multicollinearity among the explanatory variables, we calculated the Variance Inflation Factor (VIF). The VIF values were examined to detect high collinearity between variables.

The illiterate rate, uneducated rate, total population, and indigenous rate variables all had VIF scores above the 2.5 threshold used, which indicates considerable collinearity with other variables.[15]

Once removing variables with high VIF scores, variables "rate_60,"(provinces ages over 60 rate) "UNEMP_rate," "mestizo_rate," "white_rate," and "black_rate" showed acceptable VIF values, indicating no severe multicollinearity issues (VIF values: rate_60 = 1.486, UNEMP_rate = 2.086, mestizo_rate = 1.840, white_rate = 1.559, black_rate = 1.944).

Our LM equation:

$$EDF = \beta_0 + \beta_1 \cdot rate_{60} + \beta_2 \cdot UNEMP_{rate} + \beta_3 \cdot mestizo_{rate} + \beta_4 \cdot white_{rate} + \beta_5 \cdot black_{rate} + \epsilon$$

represents a linear regression model where the dependent variable, EDF (Excess Death Factor), is modeled as a linear combination of several explanatory variables. The explanatory variables include `rate_60`, `UNEMP_rate`, `mestizo_rate`, `white_rate`, and `black_rate`. The coefficients β_1, β_2, β_3, β_4, and β_5 represent the estimated effects or contributions of each explanatory variable on the EDF in each province.

The intercept term, `\(\beta_0\)`, represents the estimated value of EDF when all the explanatory variables are zero. It captures the baseline level of EDF that is not influenced by the other variables.

The error term is denoted as `\(\epsilon\)`, which represents the discrepancy between the observed values of the dependent variable (EDF) and the values predicted by the model based on the estimated coefficients and explanatory variables. It captures the unexplained or random variation in EDF that is not accounted for by the explanatory variables.

AIC Values

	Models
	AIC Value
LM	-0.87
SAR	-6.52
CAR	-3.12

Figure 5 : Monte-Carlo Simulation of Moran's I

Furthermore, I conducted a Moran's I test to assess the presence of spatial autocorrelation in the residuals of the LM model. The Moran's I test did not reveal significant spatial autocorrelation in the residuals(p > 0.87), suggesting that the model adequately captured the spatial structure of the data (Figure 5). As mentioned, the absense of spatial autocorrelation entails that a linear regression method could be used instead of

Next, I will present the results of the SAR and CAR models, compare their goodness of fit using the Akaike Information Criterion (AIC), and provide interpretation and insights regarding the significance of socioeconomic factors on EDF in Ecuadorian provinces during the COVID-19 period.

AIC Values

	Models

	AIC Value
LM	-0.87
SAR	-6.52
CAR	-3.12

Figure 6 : AIC Values for LM, SAR, and CAR Models

Lambda Values

	Models
	Lambda
SAR	-0.31042
CAR	-0.35001

Figure 7 : Lambda Values for CAR and SAR Models

After creating SAR and CAR models, it is evident that the better model, again where a lower AIC suggests that the model provides a better trade-off between goodness of fit and model complexity, is the SAR model with an AIC of -6.52 (Figure 6). The estimated lambda value for the SAR model is -0.31042, and represents the spatial autocorrelation parameter (Figure 7). Here, a negative lambda value indicates negative spatial autocorrelation, and the SAR model's lambda of -0.31042 suggests a slightly weaker negative spatial autocorrelation compared to the CAR model.

SAR Model

Characteristic	Beta	95% CI	p-value
(Intercept)	0.82	0.56, 1.1	**<0.001**
rate_60	1.9	-0.57, 4.3	0.13
UNEMP_rate	55	44, 65	**<0.001**
mestizo_rate	-0.22	-0.64, 0.20	0.3
white_rate	-4.3	-8.2, -0.43	**0.030**
black_rate	-12	-15, -8.8	**<0.001**
lambda	-0.31	-0.39, -0.23	**<0.001**

$$Y = X'\beta + \lambda W(Y - X'\beta) + \varepsilon$$

where: Y is the dependent variable (EDF in this case), X is the matrix of independent variables (including rate_60, UNEMP_rate, mestizo_rate, white_rate, and black_rate), β is the vector of coefficients, λ is the spatial autoregressive coefficient, W is the spatial weights matrix, ε is the error term.

Figure 8 : Model summary for the SAR model

Extracting the SAR model, we find that Unemployment rate, Black population percentage, and White population percentage are significant factors in predicting EDF in Ecuadorian provinces (Figure 8).

The unemployment rate (UNEMP_rate) is a significant predictor of the Excess Death Factor (EDF) in Ecuadorian provinces during COVID-19. For each one-unit increase in the unemployment rate, the expected value of EDF increases by 54.8, on average, while holding other variables constant. This suggests that higher levels of unemployment are associated with a higher EDF, indicating a potential correlation between unemployment and the severity of the COVID-19 impact.

The difference in coefficient magnitude between the rate of White individuals (white_rate) and the rate of Black individuals (black_rate) is notable in the SAR model.

The coefficient for white_rate (-4.32329) is negative and indicates that a higher proportion of White individuals in a province is associated with a decrease in the expected value of the Excess Death Factor (EDF). In other words, an increase in the rate of White individuals is linked to a potentially lower vulnerability or protective effect against the impact of COVID-19.

On the other hand, the coefficient for black_rate (-12.01226) is also negative but larger in magnitude than the coefficient for white_rate. This implies that a higher proportion of Black individuals in a province is associated with an even greater decrease in the expected value of EDF. This suggests that the rate of Black individuals may have a stronger protective effect or lower vulnerability to the impact of COVID-19 compared to the rate of White individuals.

The difference in coefficient magnitude between white_rate and black_rate suggests that the rate of Black individuals may have a more pronounced influence on the EDF in Ecuadorian provinces during the COVID-19 pandemic. This could be indicative of specific demographic, social, or healthcare-related factors that contribute to the observed association. Further research and exploration are necessary to better understand the underlying reasons for this difference and its implications for public health interventions and policy decisions.

Discussion

The significant findings of this analysis emphasize the crucial role of socioeconomic factors in understanding the variability in the Excess Death Factor (EDF) across Ecuadorian provinces during the COVID-19 pandemic. The unemployment rate, as well as the rates of White and Black individuals, have emerged as significant predictors of the severity of the pandemic's impact. These findings underscore the importance of considering broader contextual factors when examining the distribution of COVID-19 effects within a country.

It is crucial to acknowledge certain limitations and caveats associated with this analysis. Firstly, the study did not encompass an investigation of the Galapagos Islands, which present a unique geographic and demographic context within Ecuador. Future research should explore the specific dynamics of the Galapagos Islands and their potential influence on the EDF.

Moreover, it is worth noting that the Moran's I test did not reveal significant spatial autocorrelation. While this suggests that the spatial dependence among the EDF values is not pronounced, further investigation is needed to fully understand the spatial patterns and potential clustering of the EDF.

Additionally, although health site data was available, it was not incorporated into the analysis. Integrating spatial data on health sites by province could provide valuable insights into the distribution of healthcare resources and infrastructure, income disparities, job security, and other socioeconomic factors. This comprehensive approach would contribute to a more nuanced understanding of the healthcare needs and potential vulnerabilities associated with COVID-19 in Ecuador.

Further research is warranted to delve into the underlying mechanisms and implications of the identified associations. Exploring the specific pathways through which unemployment rates and the rates of White and Black individuals influence the EDF could shed light on the complex interactions between socioeconomic factors and COVID-19 outcomes.

In conclusion, this study highlights the significance of socioeconomic factors in shaping the variation in the EDF across Ecuadorian provinces during the COVID-19 pandemic. By considering the unemployment rate and the rates of White and Black individuals, we gain valuable insights into the multifaceted nature of the pandemic's impact. Future research should expand the analysis to include the Galapagos Islands and incorporate spatial data on health sites and additional socioeconomic indicators. This would provide a more comprehensive understanding of the factors influencing the severity of the pandemic and inform strategies for effective public health responses in Ecuador.

References

[1] CARE. (2020, July 17). *Uncollected Bodies Lie for Days in the Streets of Guayaquil, Ecuador, the Emerging Epicenter of the Coronavirus in Latin America.* Retrieved from www.care.org/news-and-stories/media-coverage/uncollected-bodies-lie-for-days-in-the-streets-of-guayaquil-ecuador-the-emerging-epicenter-of-the-coronavirus-in-latin-america/

[2] Cuéllar, L., Torres, I., Romero-Severson, E., Mahesh, R., Ortega, N., Pungitore, S., Hengartner, N., et al. (2021, March 1). *Excess Deaths Reveal the True Spatial, Temporal, and Demographic Impact of Covid-19 on Mortality in Ecuador. medRxiv: The Preprint Server for Health Sciences.* Retrieved from www.ncbi.nlm.nih.gov/pmc/articles/PMC7941667/table/T1/?report=objectonl

[3] Woolf, S. H., Chapman, D. A., Sabo, R. T., Weinberger, D. M., Hill, L., & Taylor, D. D. H. (2020). *Excess deaths from COVID-19 and other causes, March-July 2020. JAMA, 324*(1562-1564).

[4] Weinberger, D. M., Chen, J., Cohen, T., et al. (2020). *Estimation of excess deaths associated with the COVID-19 pandemic in the United States, March to May 2020. JAMA Internal Medicine, 180*(1336-1344).

[5] Bilinski, A., & Emanuel, E. J. (2020). *COVID-19 and excess all-cause mortality in the US and 18 comparison countries. JAMA, 324*(2100-2102).

[6] Rossen, L. M., Branum, A. M., Ahmad, F. B., Sutton, P., & Anderson, R. N. (2020). *Excess deaths associated with COVID-19, by age and race and ethnicity - United States, January 26-October 3, 2020. MMWR. Morbidity and Mortality Weekly Report, 69*(1522-1527).

[7] Our World in Data. (2020). *Coronavirus Pandemic (COVID-19).* Retrieved May 31, 2021, from ourworldindata.org/coronavirus

[8] Andrab. (2023). *Andrab/Ecuacovid: Datos Sin Procesar Extraído, Limpiado, Y Normalizado de Los Informes de La Situación Nacional Frente a La Emergencia Sanitaria SARS-Cov2 (COVID-19) de SNGRE, MSP, Registro Civil, E INEC* [Raw data extracted, cleaned, and normalized from national situation reports regarding the health emergency SARS-CoV-2 (COVID-19) from SNGRE, MSP, Civil Registry, and INEC]. Retrieved June 5, 2023, from GitHub: github.com/andrab/ecuacovid/master/datos_crudos/muertes/por_fecha

[9] O'Neill, A. (2023, May 16). Ecuador - Age Structure 2011-2021. *Statista*. Retrieved from www.statista.com/statistics/451305/age-structure-in-ecuador/

[10] Sies. "Banco de Datos Abiertos (INEC Open Data Bank)." *Banco de Información*. Retrieved from aplicaciones3.ecuadorencifras.gob.ec/BIINEC-war/

[11] "Data Cart." *IPUMS*. Retrieved June 5, 2023, from international.ipums.org/international-action/variables/group

[12] Bivand, R., Pebesma, E., & Gómez-Rubio, V. (2013). *Applied Spatial Data Analysis with R*. Springer. Section 9, p 263-314.

[13] Negahban, N. (2020, July 22). 5 Benefits of Geospatial Visualization When Analyzing IOT Data. *IoT Agenda*. Retrieved from www.techtarget.com/iotagenda/blog/IoT-Agenda/5-benefits-of-geospatial-visualization-when-analyzing-IoT-data

[14] Ishikawa, T., & Newcombe, N. S. (2021). Why Spatial Is Special in Education, Learning, and Everyday Activities. *Cognitive Research: Principles and Implications*, 6(1). doi: https://doi.org/10.1186/s41235-021-00274-5

[15] Sawe, B. E. (2018, June 7). Ethnic Groups of Ecuador. *WorldAtlas*. Retrieved from www.worldatlas.com/articles/ethnic-groups-of-ecuador.html

[16] Choueiry, G. (2020, June 1). What is an Acceptable Value for VIF? (With References). *QUANTIFYING HEALTH*. Retrieved from quantifyinghealth.com/vif-threshold/