Association of Sociodemographic Factors with the Evolution of COVID 19 Infections: Ecuadorian Case Study

L. M. Armijos-Toro, Sergio Castillo-Páez, and Paúl Medina-Vásquez

Universidad de las Fuerzas Armadas ESPE, Sangolquí, Ecuador
sacastillo@espe.edu.ec

Abstract. This study addresses the relationship between different levels of COVID 19 infection in the provinces of Ecuador, with their sociodemographic characteristics. For this purpose, a panel data model is used to identify the individual fixed effects of each province with respect to the total number of reported infections. These fixed effects are then related, by using a multiple linear regression model with provincial dummy variables derived from a principal component analysis. For the latter one, we used different provincial information related to the distribution of its population by income, economic activity, access to basic services, education, among other factors. The results obtained show an important relationship between the total number of people infected with these variables, especially with those provinces where the level of infection is higher.

Keywords: Fixed effects models · COVID 19 · Sociodemographic variables

1 Introduction

Since the beginning of reports of contagion in Ecuador, public health authorities have recorded the behavior of the virus in this country in order to establish action plans to stop the advance of COVID 19 in the population. The strategies used by the government involved various prevention actions, ranging from a nationwide quarantine to the current provincial and cantonal restriction semaphores, among other political and economic decisions [1] which have been intensifying and modulating throughout this period. However, this study seeks to analyze the evolution of total infections by province from the detection of patient zero to the start of semaphore restrictions, as shown in Fig. 1.

According to Djalante et al. [6] strategies based on disaster resilience can contribute to the management of this pandemic, and one of these strategies involves the use of scientific evidence for health management; indeed, resorting to scientific methodology to identify means, resources, places and characteristics of the population becomes indispensable to protect the population and prevent the uncontrolled spread of the virus.
In this line, epidemiology through mathematical models allows studying the evolution of the contagion. However, as [5] states, basic epidemiological models usually assume, among other things, that both the speed of contact and the probability of transmission are constant, and that infected and uninfected individuals coincide at random. This author, in his study, recognizes that the behavior of individuals when faced with an infectious disease is dynamic. In the case of non-infected individuals, they determine their preventive decisions based on the costs of prevention versus the benefits of not contracting the disease in the future. In the case of infected individuals, this author indicates that their behavior is not necessarily altruistic in avoiding the spread of the contagion, but depends on the circumstances. Other studies even determine various social and economic factors that affect the levels of contagion of a disease. For example, Dupas [4] establishes that financial limitations, deficient provision of preventive services, lack of information or incorrect information, and miseducation, among other factors, affect the demand for health prevention by the vulnerable population.

Specifically with respect to COVID 19, there is evidence of a possible relationship between sociodemographic variables of the population and levels of contagion. For example, it is well known that among the risk factors for COVID 19 infection is obesity (see e.g. [7]). In his study, [8] was able to establish the relationship of these factors with the so-called metabolic syndrome ¹ with economic variables, such as national income and the degree of urbanization of the population, among others. Recent studies conducted by [10] have used regression techniques to relate certain socioeconomic factors and political characteristics of the population with the degree of social distancing due to COVID 19. These authors conclude that sectors of the population with low socioeconomic levels are correlated with reduced social distancing. Moreover, [9] uses several linear models with panel data to verify the impact of social and economic factors on coronavirus transmission in China. Their results indicated that cities with higher

---

¹ Metabolic syndrome refers to the presence of three or more of the following conditions: excess abdominal fat; high blood sugar level; hypertension; high triglycerides, and high total cholesterol [11].
GDP per capita had higher infection rates, this due to the high level of social interactions and economic activity.

In the present work, the relationship between certain socioeconomic factors and the behavior of coronavirus infection at the provincial level in Ecuador was studied. To do so, we used a linear autoregressive regression model with fixed effects panel data to estimate the daily total of infections per province. Subsequently, the results of this model were related, by means of multiple linear regression, to the main components derived from selected sociodemographic variables. The data and variables used, as well as the details of the statistical models implemented are presented in Sect. 2. The results of these models are found in Sect. 3 and the main findings of this work are presented in Sect. 4.

2 Data and Models Analysed

2.1 Information Sources

For the present work, information about the number of people confirmed with COVID-19 by province, in Ecuador, was obtained from official sources; in particular, from the National Secretariat of Risk and Emergency Management [1] through its situation reports. The data analysed corresponds to the total daily number of positive COVID-19 tests (TDNP) reported by province from the date of first registration, since March 13, to April 9 2020. Subsequent records were not included, due to April 10 significant sporadic increases in these records were evident, as shown in Fig. 2.

![Fig. 2. Diagnostic tests per day - Ecuador](image-url)

During the study period, 9,639 test results were obtained, of which 4,965 were positive diagnoses, distributed in the following provinces: Guayas, Pichincha, Los Ríos, Manabí, Sucumbíos, Azuay, El Oro, Morona Santiago, Santa
Elena, Bolívar, Cañar, Chimborazo, Imbabura, Loja and Santo Domingo de los Tsáchilas. Figure 3 shows the number of confirmed positive cases of COVID-19 as of April 9, by province.

![Number of confirmed cases of COVID-19 by province under analysis up to 9 April 2020 - Ecuador](image)

**Fig. 3.** Number of confirmed cases of COVID-19 by province under analysis up to 9 April 2020 - Ecuador

With regard to the sociodemographic variables of the provinces used in this work, the National Urban Employment and Unemployment Survey - accumulated [2] for the year 2018 has been used. Specifically, the following variables were analysed: Percentage of overcrowded households (POH), Percentage of households with access to basic services (PHS), Income Poverty (IP), Gini Inequality (G), Percentage of urban population in the province (PPU), Percentage of informal sector in the province (IPP), Percentage of population with less than full employment (LE), Average years of schooling (YS) and Percentage of population between 15 and 50 years of age (PPY) in the province.

It should be noted that the variable (G) was multiplied by 100 in order to make its scale similar to that of the other data. Additionally, information from the laboratory of labor dynamics of INEC [3] was included, related to the percentage distribution of the number of commercial establishments per province with respect to the total number of commercial establishments in the country (NCP), in order to have data related to the provincial economic activity.

### 2.2 Quantitative Methods and Models Used

In a first instance, it was evaluated whether the daily total of infections has a different effect per province. For this, a panel data model was used, using daily positive test records (TDNP) by each province. Initially, a fixed-effect static model was considered, as described in Eq. (1)

$$TDNP_{it} = \beta(TDNP_{i(t-1)}) + \alpha_i + \epsilon_{it},$$  

(1)
where $TDNP_{it}$ is the total number of tests with confirmed cases of COVID-19 from the province $i$ in the day $t$; $TDNP_{i(t-1)}$ is its corresponding 1-day delay; $\beta$ is the effect of the autoregressive process, while $\alpha_i$ is the individual effect of each province $i$ on the value of $TDNP_{i,t}$, and $\epsilon_{it}$ is the term of the error [12].

Under this model, $\beta$ allows to estimate the rate of evolution of the disease’s spread in general. On the other hand, if there is any significant difference in the total daily affirmative tests $TDNP_{i,t}$ between the provinces considered in the study, this effect is identified in the terms $\alpha_i$ of the Eq. (1).

The next step was to check whether the fixed effects of each province $\alpha_i$ are statistically related to the variables of interest considered, which are POH, PHS, IP, G, PPU, IPP, LE, YS, PPY and PCB. This process was divided into two stages: first, the number of sociodemographic variables was reduced using principal component analysis. This technique allows us to obtain a new reduced set of dummy variables $Z_1, Z_2, Z_3, \ldots, Z_r$, not correlated to each other, so that these $r \leq 10$ variables explain a large percentage of the variance of the data. Secondly, once the variables $Z_k$ have been identified with $k = 1, 2, 3, \ldots, r$, a linear regression model is constructed, considering the estimated fixed effects of each province $\hat{\alpha}_i$ with the $r$ main components obtained from the sociodemographic variables considered, obtaining the model shown in the Eq. (2)

$$\hat{\alpha}_i = \mu + \lambda_1 Z_1 + \lambda_2 Z_2 + \ldots + \lambda_r Z_r.$$ (2)

Once model (2) has been estimated, diagnostic and validity tests were performed to verify whether the correlations between the estimated fixed effects, individual for each province on $TDNP_{i,t}$, and the main components obtained from the sociodemographic variables, represented by the coefficients $\lambda_k$, are statistically significant.

3 Results

3.1 Panel Data Model

Considering that, in general, a panel data model combines information of cross-sectional variables in different time intervals, which in our case are the records of $TDNP_{it}$ for each province $i$ in the day $t$, due to daily records vary in each province, a unbalanced panel data model was considered to estimate.

The results obtained, model (1), show a statistically significant $\hat{\beta} = 1.054 \ (t = 202.59, P - Value < 2.2e-16)$. In addition, the correlation and overall significance coefficients were also considered adequate, since $(R^2 = 0.9919$, adjusted $R^2 = 0.9915$, $P - Value(F_{test}) < 2.22e - 16)$; and, for individual $\alpha_i$ effects by province, the results of their estimation are shown in Table 1.

To check whether the fixed-effect model is suitable, several verification and diagnostic tests were performed. The results are presented in Table 2.

From Table 2 it can be concluded that: i) the $\alpha_i$ effects are non-null and non-random; ii) there are no fixed effects related to time; and, iii) the data used

\[2\] A detailed explanation of the procedure used can be seen in [13].
Table 1. Estimates of individual fixed effects by province

| Province | α  |
|----------|----|
| Guayas   | 60.43421 |
| Pichincha | 9.70178  |
| Manabí   | 4.08408  |
| Los Ríos | 3.16396  |
| El Oro   | 2.56321  |

| Province | α  |
|----------|----|
| Santa Elena | 1.83335 |
| Cañar | 1.78121  |
| Azuay | 1.75986  |
| Chimbéorazo | 1.41254 |
| Santo Domingo | 1.18817 |

| Province | α  |
|----------|----|
| Loja     | 1.22449  |
| Imbabura | 0.49337  |
| Bolivar  | 0.44067  |
| Sucumbíos | 0.37203  |
| Morona Santiago | 0.32224 |

Table 2. Panel data model validation tests

| Test                        | H₀           | P Value   |
|-----------------------------|--------------|-----------|
| Test F for individual effects | Null fixed effect | 4.372e−06 |
| Test of Hausman             | Assume random effects | 1.684e−14 |
| Test F (Lagrange)           | Null temporal fixed effects | 0.2893 |
| Test Dickey-Fuller          | Non-stationary data | 0.02179 |

are stationary. Therefore, we can assume that the \(\alpha_i\) estimates, presented in the Table 1, represent the individual effect of each province on the \(TDNP\) variable.

3.2 Principal Component Analysis

The analysis of principal components allowed the reduction of the information of the sociodemographic variables, obtaining at the same time \(Z_k (k = 1, \ldots, r)\) uncorrelated dummy variables. Prior to the application of this technique, it was verified by means of the Bartlet test \((\text{Chi-Square} = 147.248, P - \text{Value} < 8.63e−13)\) that the application of this technique is adequate. Then, it was necessary to select a small number of main components for the analysis. From the results obtained, it can be seen that using three, denoted by, \(Z_1, Z_2\) and \(Z_3\) explains 86% of the variance of the original data set (60.2%, 13.2% and 12.6%, respectively).

In order to interpret the possible relationship of each of the sociodemographic variables, with the selected components \(Z_1, Z_2\) and \(Z_3\), the graphs shown in Fig. 4 were constructed, in which the contribution of each sociodemographic variable for each component is evident.

From this last Figure, it can be seen that the first component \(Z_1\) has a high positive correlation with the variables: \(PCB, PPY, YS\) and \(PPU\); while negatively with: \(LE, IP\) and \(IPP\). The component \(Z_2\) has a higher positive correlation with the variables: \(PHS\) and \(G\); while negatively with \(POH\). Finally, the component \(Z_3\) has a higher correlation with: \(G\) followed by \(PPY, PCB\) and \(IP\). By way of summary, the main components can be described as indicated in the Table 3:
Fig. 4. Contribution charts of sociodemographic variables for the components, in (a) $Z_1$ vs. $Z_2$ and in (b) $Z_2$ vs. $Z_3$.

Table 3. Characterization of components $Z$

| Component | Characterization                                      |
|-----------|-------------------------------------------------------|
| $Z_1$     | Demographic and economic characteristics of the province |
| $Z_2$     | Household infrastructure characteristics              |
| $Z_3$     | Characteristic of inequality in the province          |

On the other hand, in Fig. 5 you can see how the provinces are related according to two selected principal components, in particular, $Z_1$ vs. $Z_2$. It is interesting to note that the provinces of Guayas and Pichincha, which have the highest number of infections, are those furthest away from the data cloud, although in different quadrants.

3.3 Linear Regression Model

Once the dummy variables $Z_1$, $Z_2$ and $Z_3$ had been selected, their statistical correlation was established with the estimated individual infection effects, by province, given by the terms $\hat{\alpha}_i$. Since these estimates do not present a symmetrical behavior, it was necessary to do a logarithmic transformation $\log(\hat{\alpha})$ before its linear modeling. The results of this transformation are shown in Fig. 6.
Using the latter transformation, the model was redefined (2) to estimate a linear regression model, based on the components \( Z_1, Z_2 \) and \( Z_3 \), as described in Eq. (3)

\[
\log(\hat{\alpha}) = \mu + \lambda_1 Z_1 + \lambda_2 Z_2 + \lambda_3 Z_3. \tag{3}
\]

However, although the fit of this model is adequate \((R^2 = 0.7694, R^2 \text{ fit } = 0.7065)\), the \( \lambda_3 \) coefficient is not significant \((t = 1.6654, P - \text{Value} = 0.1264)\); therefore, the model was modified again, removing the \( Z_3 \) component from the Eq. (3). The results of the estimation of the last-mentioned model are shown in Table 4:

**Table 4.** Linear regression model of \( \log(\hat{\alpha}) \), with respect to the components \( Z_1 \) y \( Z_2 \)

| Coefficients | Estimator | St. Error | t     | P Value |
|--------------|-----------|-----------|-------|---------|
| \( \mu \)    | 0.57995   | 0.20303   | 2.856 | 0.014450|
| \( \lambda_1 \) | 0.42629   | 0.08563   | 4.978 | 0.000321|
| \( \lambda_2 \) | -0.40444  | 0.18296   | -2.211| 0.047233|

The goodness-of-fit of this model is considered adequate, as \((R^2 = 0.712, R^2 \text{ adjusted } = 0.664)\) and, in addition, the overall significance is not impaired \((F = \ldots \text{ continues})\).
14.84, \( P - \text{Value} = 0.0005703 \). Likewise, performing the validity and diagnostic tests, it is verified that the model complies with all the basic assumptions, as shown in Table 5:

| Test                  | \( H_0 \)          | P Value |
|-----------------------|---------------------|---------|
| Jarque - Bera         | Normality of errors | 0.066   |
| Breusch - Pagan       | Homocedasticity     | 0.7693  |
| Durbi - Watson        | Null autocorrelation| 0.388   |

4 Discussion

The results obtained by the modified panel data model coincide with that presented in Sect. 1, in that the levels of infection do not necessarily have to be similar between provinces, but rather there are local factors that influence the total number of people infected by COVID 19. Specifically, it is observed that the provinces of Guayas and Pichincha, where the highest level of infection is registered, present higher individual effects than the rest of the provinces (see Table 1). This behaviour is corroborated by the analysis that these provinces present different levels of association with the principal components, unlike the rest of the provinces (see Fig. 5).

Besides, the analysis of main components allowed the identification of the most representative factors at the provincial level, being those related to the economic activity and the infrastructure of the households in each province (see Table 3).

Finally, the linear regression model made it possible to explain the relationship between provincial fixed effects and these representative factors. There is a positive relationship between these fixed effects and the variables related to the first principal component (\( \lambda_1 > 0 \) in Table 4); therefore, it is to be expected that in those provinces that have a higher economic activity, the fixed effect of that province will also be higher, and so will its level of contagion. A similar relationship, but in an inversely proportional sense, can be deduced between the infrastructure characteristics of households and the individual effect on the level of infection in each province (\( \lambda_2 < 0 \) in Table 4). This can be clearly evidenced by the locations of Guayas and Pichincha in Fig. 5, because although their level of economic activity is similar, Pichincha is somewhat more, we assume that since it is the seat of government and is the place where many companies have their fiscal domicile, unlike Guayas, that their activity is highly commercial; the level of infrastructure of the homes is very different, and it is therefore highly variable. Here three aspects can be considered: i) that in reality the infrastructure between both provinces at the housing level is very different (better in one
than in the other), ii) the geographic location and climate make the difference or iii) a combination of the two previous. For these considerations it is important to note that with the exception of the El Oro province, all provinces in the coastal region are in the negative part of the second principal component (see Fig. 5).

We consider that this study, although based on official data provided by state authorities, with the limitations of the case, allows to establish significantly the relationship between the sociodemographic conditions of each province and the level of contagion, given that the proposed models is relational, it does not seek to be a predictive model, so when analyzing the data to date (see Fig. 7) it is confirmed that the evolution of infections is greater in the provinces of Guayas and Pichincha, as expected. Its results would allow control agencies to establish appropriate health prevention strategies differentiated based on specific characteristics of the population and the province.

![Fig. 7. Series of infections in the provinces of Guayas and Pichincha](image)

**Acknowledgements.** The authors would like to thank the Department of Exact Sciences of Universidad de las Fuerzas Armadas ESPE, for the support provided for the development of this study.

**References**

1. Secretaría Nacional de Gestión de Riesgos y Emergencias. Informes de Situación e infografías. [https://www.gestionderiesgos.gob.ec](https://www.gestionderiesgos.gob.ec)

2. Instituto Nacional de Estadística y Censos. Encuesta acumulada ENEMDU (2020). [https://www.ecuadorencifras.gob.ec/documentos/web-inec/Sitios/ENEMDU_ACUMULADA](https://www.ecuadorencifras.gob.ec/documentos/web-inec/Sitios/ENEMDU_ACUMULADA)

3. Instituto Nacional de Estadística y Censos. Laboratorio de dinámica laboral y empresarial. [https://www.ecuadorencifras.gob.ec](https://www.ecuadorencifras.gob.ec)
4. Dupas, P.: Health behavior in developing countries. Annu. Rev. Econ. 3(1), 425–449 (2011)
5. Hauck, K.: The Economics of Infectious Diseases, May 2018. https://doi.org/10.1093/acrefore/9780190625979.013.251
6. Djalante, R., Shaw, R., DeWit, A.: Building resilience against biological hazards and pandemics: COVID-19 and its implications for the Sendai Framework. Prog. Disaster Sci. 6, 100080 (2020). https://doi.org/10.1016/j.pdisas.2020.100080
7. Dietz, W., Santos-Burga, C.: Obesity and its implications for COVID-19 mortality. Obesity 28(3), 1005–1005 (2020)
8. Danaei, G., Singh, G.M., Paciorek, C.J., Lin, J.K., Cowan, M.J., Finucane, M.M., Farzadfar, F., Stevens, G.A., Riley, L.M., Lu, Y., Rao, M., Ezzati, M.: The global cardiovascular risk transition: associations of four metabolic risk factors with national income, urbanization, and western diet in 1980 and 2008. Circulation 127(14), 1493–1502 (2013). https://doi.org/10.1161/CIRCULATIONAHA.113.001470
9. Qiu, Y., Chen, X., Shi, W.: Impacts of social and economic factors on the transmission of coronavirus disease 2019 (COVID-19) in China. J. Popul. Econ. 2019(13165) (2020). https://doi.org/10.1007/s00148-020-00778-2
10. Kavanagh, N.M., Goel, R.R., Venkataramani, A.S.: Association of County-Level Socioeconomic and Political Characteristics with Engagement in Social Distancing for COVID-19, medRxiv (2020)
11. Múnera, M., Escobar, S.: Carta de Laboratorio Clínico ). Obras de la Congregación Mariana. Medellín, Colombia (2007)
12. Arellano, M., Bover, O., et al.: La econometría de datos de panel. Investigaciones económicas 14(1), 3–45 (1990)
13. Härdle W.K., Simar, L.: Principal components analysis. In: Applied Multivariate Statistical Analysis, pp. 269–305. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-17229-8_10