Using Functional Data Analysis to determine the relationship between socioeconomic and demographic factors and the infection rates of Covid-19 in the different communes of Santiago de Chile

Abelino Sepulveda Estrada [1]

Advisors:
Nicolas Alberto Moreno Reyes[2]

Research practice 1
Final Report
Mathematical Engineering
Department of Mathematical Sciences
School of Sciences
Universidad EAFIT

May 2022

[1]Mathematical Engineering student, Universidad EAFIT. asepulvede@eafit.edu.co
[2]Professor in Department of Mathematics Sciences, Universidad EAFIT. namorenor@eafit.edu.co

Abstract

The COVID-19 pandemic has so far caused huge negative impacts on different areas all over the world, and Santiago de Chile was no exception. In this paper we use functional data analysis to study the behavior of the infection rates of each commune of this city. In addition, we perform a permutation test and a semiparametric permutation test to find if there is a difference between the infection rates of each commune. Finally, we perform a linear regression to determine if there are socioeconomic and demographic factors that influence this rate.

Keywords: Functional Data, Permutation Test, Correlation, Multiple Linear Regression, Covid-19

1 Introduction

In December 2019, a cluster of severe pneumonia cases of unknown cause was reported in Wuhan, Hubei province, China (Cucinotta & Vanelli, 2020), which was later discovered to be about severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) that is a viral respiratory disease caused by a SARS-associated coronavirus that causes COVID-19 (coronavirus disease 2019). A few months later, on March 11, 2020, the World Health Organization (WHO) declared the COVID-19 a global pandemic (WHO *et al.*, 2020). This virulent disease has caused around 395.000.000 confirmed and 5.740.000 death cases around the world.

The first control strategies for this disease were preventive measures such as quarantines, use of face masks, hand hygiene and social distancing (Chinazzi *et al.*, 2020). However, in populations with precarious situations, the spread of the disease could not be controlled, as in the case of Santiago de Chile, which has a socially and economically segregated population. In this city, the mobility and physical distancing restrictions became almost null for the most socioeconomically vulnerable communes, causing high rates of infection and mortality (Mena *et al.*, 2021).

Vaccines with inactivated or attenuated viruses, protein-based, with viral vectors, and RNA and DNA vaccines are currently available (WHO, 2020). In Chile, the mass vaccination program started on February 3, 2021. The strategy was gradual and chose to prioritize groups at higher risk, such as the elderly, health workers, the chronically ill, and other workers who have direct contact with people (Castro & Singer, 2021). Currently, vaccination coverage reaches 91,68% of the population with at least one dose (MINSAL, 2022).

In recent years, functional data analysis (FDA) has become a very active domain of research in statistics, for its interest and also for its applications in many contexts such as medical science, biology, and chemistry (Ramsay & Silverman, 2008). In this approach, we will use FDA tools to perform functional observations based on the graphic representation of the curves and analyze COVID-19 contagion rates and determine the influence of vaccination in the communes of Santiago de Chile. This research seeks to find out through functional data analysis if there are socioeconomic or demographic factors that influence infection rates of this disease, also to find if there are differences in the infection rates in the different communes of this city, additionally determine the influence of vaccination with the infection rates. Determining these factors is crucial, it will help develop action plans to fight disease propagation and apply preventive measures for those who were more vulnerable to the situation.

This research will serve to analyze patterns of mortality and contagion of COVID-19 and its association with mobility and socioeconomic, infrastructural and environmental covariates, especially in the city of Santiago de Chile, however it could be applied to any city, which would facilitate researchers to analyze these covariates in different cities.

Another contribution of this research is the use of different areas of functional data analysis, such as the classification of functional data Ferraty & Vieu (2006) with which we could categorize the different communes of Santiago de Chile with the rates of contagion and mortality and compare whether said categorization agrees with the results obtained by the permutation test Leng & Müller (2006). Finally, The results of the investigation could help other authors to find new covariates and their effects on the spread of COVID-19. And also it can be used as a base for future investigations.

The remainder of this paper is organized as follows. The Section 2 presents the state where studies carried out for functional data and covid-19 are investigated. Section 3 presents the techniques and methods used for the research. The results and experiments carried out are presented in section 4. Finally, in Section 5 some conclusions and future work are presented.

2 State of the art

This problem has different approaches in the literature, one of them is described by Tang *et al.* (2020) who studied COVID-19 cases in the united states through functional principal component analysis (FPCA) and explores the modes of variation of the data, and studies the canonical correlation between confirmed and death cases. The main objectives of the research were to determine if this correlation varies from state to state and if the geographical location of the infected persons was related to the spread of the disease in addition to looking for critical points in the country. Also determine if the practice of public measures helps mitigate the spread.

Boschi *et al.* (2020) investigate patterns of COVID-19 mortality across 20 Italian regions and their association with mobility, positivity, sociodemographic, infrastructural and, environmental covariates. The use functional data analysis with a focus in probabilistic k-mean with local alignment for clustering determining the sociodemographic relationship of the regions with the infection rate. On their article one of the mos interesting results is that they conclude that the levels of local mobility are strong positive predictors of mortality at their peaks of the pendemic.

A different approach has been presented by Carroll *et al.* (2020) who apply time dynamics with functional data analysis to analyze the trajectory of COVID-19 and quantify and compare the infection and death rates in 64 countries that were sampled. This framework facilitates the use of a time-varying regression to quantify the effects of demographic covariates and social mobility on duplication rates and case fatality rates as well Boschi *et al.* (2020). An interesting finding of this article is that the trajectory of a country during the first month largely determines how the transmission will develop later.

In a study conducted by Oshinubi *et al.* (2021) they use functional data analysis to model the daily cases of deaths and hospitalizations in the different departments of France. For the modeling of this study, it is taken into consideration before and after vaccination against the disease began

in this country. Finally, they use functional principal component analysis techniques and groupings were made using k-means techniques to understand the dynamics of the pandemic in the different departments.

Finally, A particular use of permutation tests was proposed by Cabassi *et al.* (2017) to generalize the metric-based permutation test for the equality of covariance proposed by Pigoli *et al.* (2014)using the non-parametric combination method of Salmaso & Pesarin (2010). The generalization of this test allows us to infer whether there is inequality of the covariance operators of the multiple groups of functional data. The null hypothesis of this test consists in assuming that there is no difference in the variance between the groups and if there is evidence to reject the null hypothesis, it helps us to identify the pairs of groups that have different covariances.

3 Methodology

Describe in this section the methods, techniques, algorithms, etc. that you have developed during the research practice.

This methodology consists of several stages: the functional construction of the data, the performing of the permutation test, and finally, with the results obtained by the permutation test, the pertinent studies are carried out to examine possible correlations.

3.1 Data Collection and pre-processing

One of the main components of this study is the collection of data. These were provided by the Ministry of Science, Technology, Knowledge, and Innovation, which had daily records of incidence rates from March 30, 2020, to May 27, 2022. The data for this study was found at the following link: `https://github.com/MinCiencia/Datos-COVID19/tree/master/output/producto18`. Once the daily records are obtained, we filter the required observations, in this case, the communes of Santiago de Chile.

3.2 Functional Data Construction

Once the data is obtained and processed, we perform the functional data construction. For this, we consider the set of discrete observations $y_{i1}, y_{i2}, ..., y_{in}$ with $i \in \{1, 2, .., 32\}$. Where each of the sets represents the incidence rate in each of the 32 communes of Santiago de Chile at a time t.

Now, Let be

$$X = x_1(t), x_2(t), ..., x_{32}(t)$$

$$x_1, x_2, ..., x_{32} \in L^2(\Omega)$$

Where $x_i(t)$ is the curve that represents the infections by COVID-19 in the commune i of the city.

Since we need to work with functions, consequently, we require a strategy for constructing functions that works with parameters that are easy to estimate and that can accommodate nearly

any curve feature, no matter how localized. We use a set of functional building blocks $\Phi_k, k = 1, ...K$ called basis functions which are combined linearly.

$$x(t) = \sum_{k=1}^{K} c_k \Phi_k(t) = c'\Phi(t)$$

Where, the parameters $c_1, c_2, ..., c_K$ are the coeficients of the expansion.

The Figure 1 shows the functional data construction of the observations for each commune.

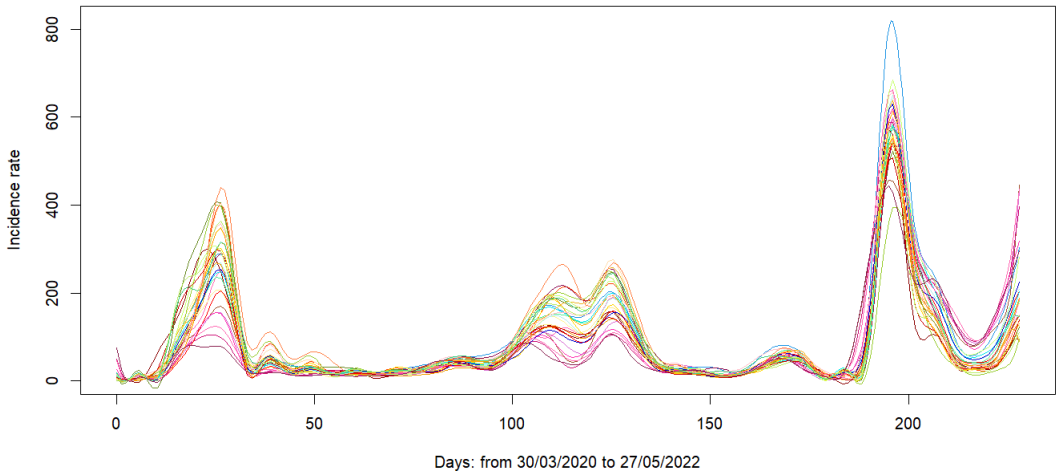

Figure 1: Functional Data for the incidence rate for each commune

3.3 Performing of the Permutation Tests

The next stage is permorfing the permutation test. For this, we consider:

3.3.1 Permutation Test to Functional Data

We use a permutation test to evaluate whether the infection rates curves for the two groups are different (T-Test). Where

$$T(t) = \frac{|\bar{x}_1(t) - \bar{x}_2(t)|}{\sqrt{\frac{1}{n_1}Var[x_1(t)] + \frac{1}{n_2}Var[x_2(t)]}}$$

We would like to test the hypothesis

$$H_0 : \bar{x_1} = \bar{x_2} \quad vs \quad H_1 : \bar{x_1} \neq \bar{x_2}$$

This test provides a sense of the difference between the infection rate between the different communes of the city.

Now, to perform this test, we need to find the infection rates of each commune and separate them into two groups. The functions that describe the behavior of the incidence rates are derived,

and we perform clustering to classify the infection rates into two groups. The Figure 2 shows the incidence rates derivative for each commune it means, the infection rates, and the Figure 3 shows the two groups of infection rates formed by clusters.

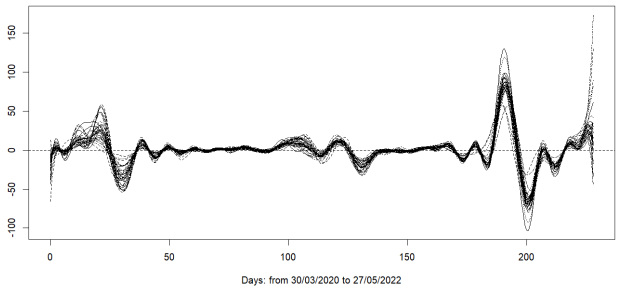

Figure 2: Incidence rates derivative: infection rates

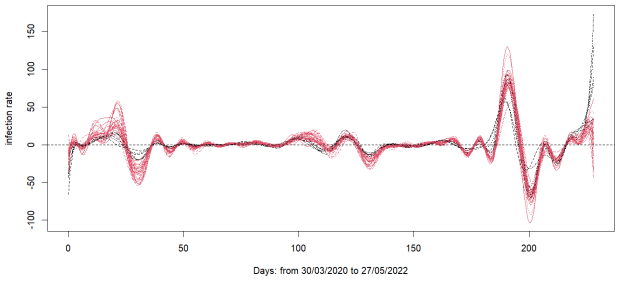

Figure 3: Clusters for the infection rates of each commune

3.3.2 Semiparametric Permutation test

In addition to the permutation test, we perform a semiparametric permutation test. Which consists of randomly multiplying with equal probability each curve with the plus or minus signs. Once these permutations have been performed, we want to test whether

$$\sum_{i=1}^{K} x_i^*(t) - \sum_{i=1}^{K} x_i(t) = 0, \qquad \text{for all t}$$

donde $x_i^*(t)$ are the permutations of the observations, and $x_i(t)$ are the observations of the infection rates.

Once these differences are calculated, we look at the p-value

$$\text{p-value} = \frac{1}{N} \sum_{j=1}^{N} I(T_j > t_{obs}$$

Where N is the number of permutations, T_j is $\sum_{i=1} K x_i^*(t)$ for all t and t_{obs} is $\sum_{i=1} K x_i^(t)$ for all t.

Finally, if the p-value gives values greater than 0.05, it indicates that the difference in the infection rates of each commune is different.

3.4 Multiple Linear Regression

Once we obtain the results of the permutation tests, we carry out correlation studies to determine if there are factors that influence the rates of Covid-19 infection. So we perform a multiple linear regression. For this we consider a linear relationship between the infection rate and some socioeconomic and demographic factors

$$y = \beta_0 + \beta_1 x_1 + ... + \beta_k x_k + \epsilon_i$$

Where y is the dependent variables, x_i are the predictors, $\beta_1, ..., \beta_k$ are the weights for the indepndent variables inside the regress equation, the β_0 is the intercept and ϵ_i is the error for each commune.

4 Results

The Figure 4 shows the perform of the permutation test with the two groups of infection rates presented in the Figure 3.

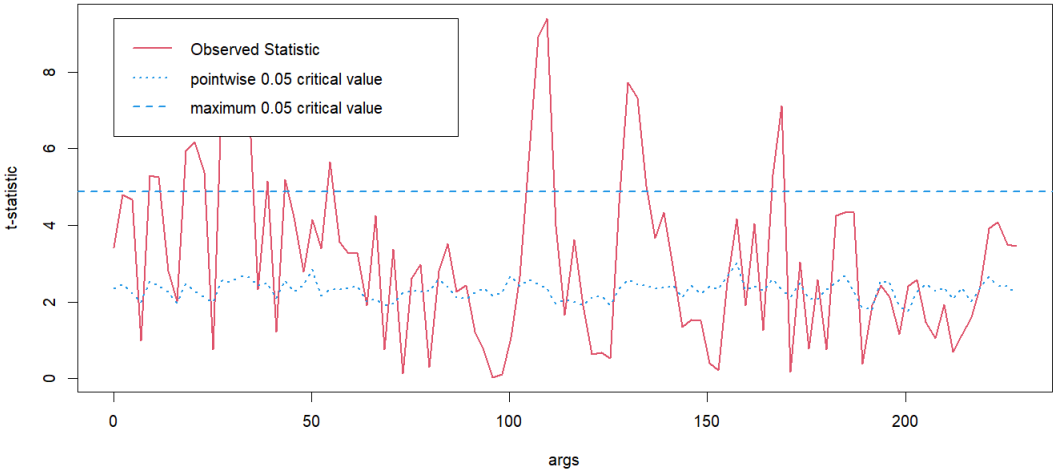

Figure 4: Permutation test for the infection rates

In this figure, we can see the observed statistic and its evolution over time. The values outside the bands represent the instants of time where the infection rates in the group are different. We can now compare the results we obtained with the semiparametric permutation test done with 100000 permutations shown in Figure 5

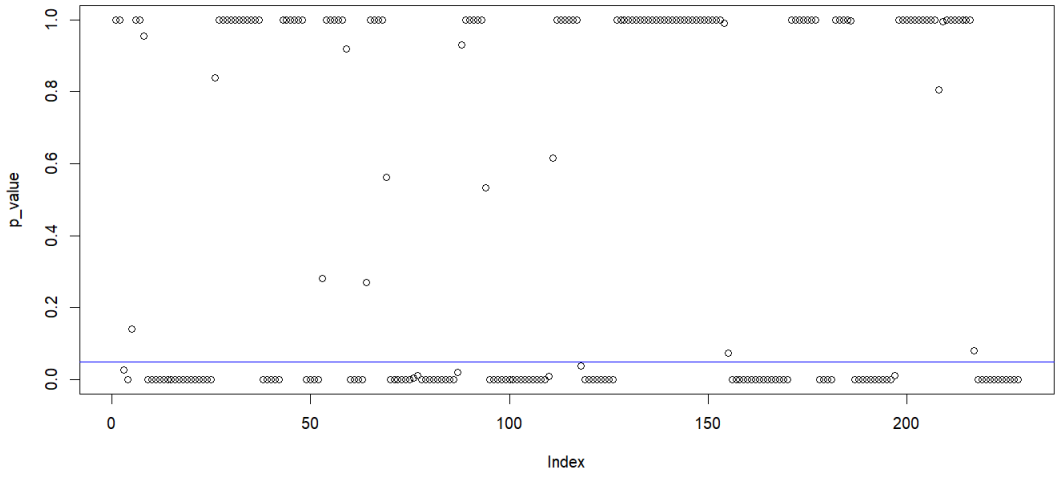

Figure 5: Semiparametric Permutation test

The results shown in the previous figure show the instants of time where the null hypothesis is rejected. Which we can see coincide with the results shown in Figure 4.

Once it is shown that there is a difference in the infection rates in the groups formed by the different communes, we begin to do the exploratory analysis to determine if there are factors that are correlated with these rates. Figure 6 shows a correlation study between the infection rates of each commune and the following factors: internal mobility index, external mobility index, socio-economic development index (IDSE), Per capita monthly income ($), poverty (%), schooling (years), if they have acceptable housing (%), if they have sewerage (%), life expectancy at birth (LEB) (years), human development index (HDI).

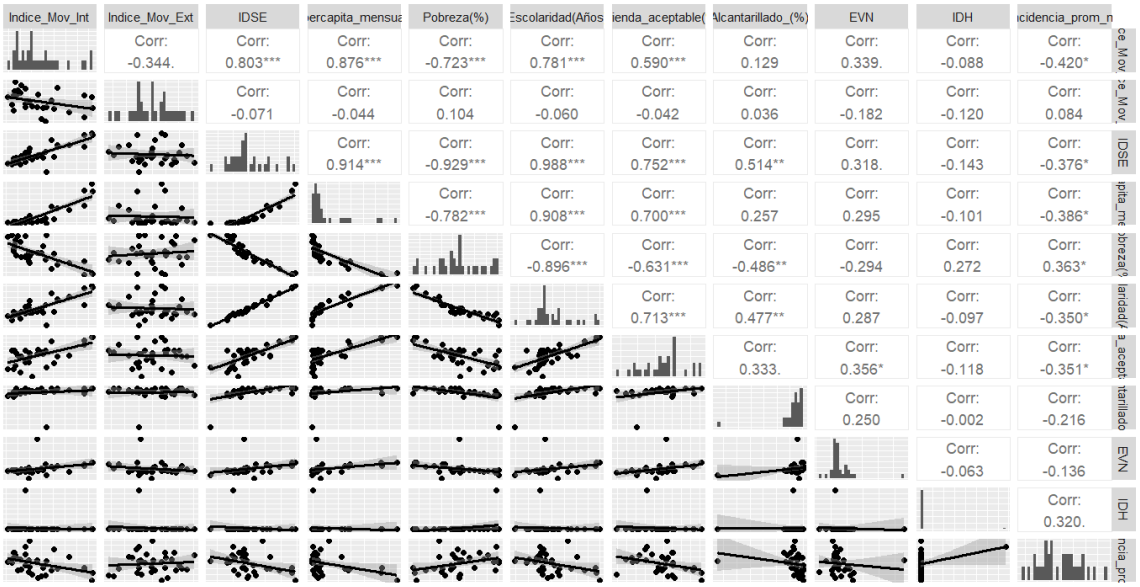

Figure 6: Correlation Matrix

The Figure 6 we can see that the factors that have a greater linear relationship with infection rates are the internal mobility index, IDSE, monthly per capita income, poverty, schooling, and if they have acceptable housing. With this correlation study we can to perform the multiple linear regression shown in the Figure 7.

```
Residuals:
    Min      1Q  Median      3Q     Max
-93.763 -50.130  -6.038  42.200 152.651

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                     1428.0071   552.5788   2.584   0.0173 *
Indice_Mov_Ext                    -5.2339    12.0575  -0.434   0.6687
Indice_Mov_Int                   -42.4498    22.0602  -1.924   0.0680 .
IDSE                            4469.7331  2425.0128   1.843   0.0795 .
`Ingreso_percapita_mensual(miles$)`   -0.3150     0.2522  -1.249   0.2254
`Pobreza(%)`                     19.9134    13.8434   1.438   0.1650
`Escolaridad(Años)`            -156.9574    98.2512  -1.598   0.1251
`Vivienda_aceptable(%)`          -10.6225     5.7543  -1.846   0.0790 .
`Alcantarillado_(%)`             -14.0744     6.2317  -2.259   0.0347 *
EVN                                3.1329     3.6502   0.858   0.4004
IDH                                2.1028     1.1797   1.782   0.0891 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 76.49 on 21 degrees of freedom
Multiple R-squared:  0.4138,    Adjusted R-squared:  0.1347
F-statistic: 1.483 on 10 and 21 DF,  p-value: 0.2142
```

Figure 7: Multiple linear regression

It is interesting to note that predictive factors such as internal mobility index, IDSE, acceptable housing, sewerage, and HDI are statistically significant in this model. Which represents how much the infection rates change for change in one unit for each factor. We can also see that there is a negative relationship between infection rates and access to sewerage if they have acceptable housing, and internal mobility index, which indicates that if these factors increase in units, these rates decrease.

5 Conclusions and future research

According to the results obtained, we can conclude that there is a difference in the infection rates of each group made up of the communes of Santiago, Chile. In addition, we can observe in what instants of the time these differences can be seen, since it does not always happen.

Also, we were able to determine the socioeconomic and demographic factors that affected the rates of covid-19 infection, and we observed if the type of relationship was positive or negative. However, due to the low adjusted R-squared and the p-value of the regression, we can affirm that there is not much reliability in the model.

Finally, we propose to carry out other correlation studies such as contingency tables or try nonlinear regression models, given that the relationship that exists between our dependent variable and the predictive factors is possibly nonlinear.

Acknowledgements

I thank Alejandro Calle for letting me attend his Functional Data Analysis master's class and for the information provided.

References

Boschi, Tobia, Di Iorio, Jacopo, Testa, Lorenzo, Cremona, Marzia A, & Chiaromonte, Francesca. 2020. The shapes of an epidemic: using Functional Data Analysis to characterize COVID-19 in Italy. *arXiv preprint arXiv:2008.04700*.

Cabassi, Alessandra, Pigoli, Davide, Secchi, Piercesare, & Carter, Patrick A. 2017. Permutation tests for the equality of covariance operators of functional data with applications to evolutionary biology. *Electronic Journal of Statistics*, **11**(2), 3815–3840.

Carroll, Cody, Bhattacharjee, Satarupa, Chen, Yaqing, Dubey, Paromita, Fan, Jianing, Gajardo, Álvaro, Zhou, Xiner, Müller, Hans-Georg, & Wang, Jane-Ling. 2020. Time dynamics of COVID-19. *Scientific reports*, **10**(1), 1–14.

Castro, Marcia C, & Singer, Burton. 2021. Prioritizing COVID-19 vaccination by age. *Proceedings of the National Academy of Sciences*, **118**(15).

Chinazzi, Matteo, Davis, Jessica T, Ajelli, Marco, Gioannini, Corrado, Litvinova, Maria, Merler, Stefano, Pastore y Piontti, Ana, Mu, Kunpeng, Rossi, Luca, Sun, Kaiyuan, *et al.* 2020. The effect of travel restrictions on the spread of the 2019 novel coronavirus (COVID-19) outbreak. *Science*, **368**(6489), 395–400.

Cucinotta, Domenico, & Vanelli, Maurizio. 2020. WHO declares COVID-19 a pandemic. *Acta Bio Medica: Atenei Parmensis*, **91**(1), 157.

Ferraty, Frédéric, & Vieu, Philippe. 2006. *Nonparametric functional data analysis: theory and practice*. Vol. 76. Springer.

Leng, Xiaoyan, & Müller, Hans-Georg. 2006. Classification using functional data analysis for temporal gene expression data. *Bioinformatics*, **22**(1), 68–76.

Mena, Gonzalo E, Martinez, Pamela P, Mahmud, Ayesha S, Marquet, Pablo A, Buckee, Caroline O, & Santillana, Mauricio. 2021. Socioeconomic status determines COVID-19 incidence and related mortality in Santiago, Chile. *Science*, **372**(6545), eabg5298.

MINSAL. 2022. Cifras Oficiales COVID-19.

Oshinubi, Kayode, Ibrahim, Firas, Rachdi, Mustapha, & Demongeot, Jacques. 2021. Functional data analysis: Transition from daily observation of COVID-19 prevalence in France to functional curves. *medRxiv*.

Pigoli, Davide, Aston, John AD, Dryden, Ian L, & Secchi, Piercesare. 2014. Distances and inference for covariance operators. *Biometrika*, **101**(2), 409–422.

Ramsay, James O, & Silverman, Bernhard W. 2008. Functional data analysis. *İnternet Adresi: http*.

Salmaso, Luigi, & Pesarin, Fortunato. 2010. *Permutation tests for complex data: theory, applications and software*. John Wiley & Sons.

Tang, Chen, Wang, Tiandong, & Zhang, Panpan. 2020. Functional data analysis: An application to COVID-19 data in the United States. *arXiv preprint arXiv:2009.08363*.

WHO, World Health Organization. 2020. *COVID-19 vaccines*.

WHO, World Health Organization, *et al.* 2020. *WHO Director-General's opening remarks at the media briefing on COVID-19-11 March 2020*.