World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Random forest using smartphone GPS in first wave of COVID-19 in the Maule region, Chile

Nicolas Ayala [1,*], Antonio Monleon-Getino [1], Jaume Canela-Soler [2] and Tomas Chadwick-Lobos [3]

[1] *Department of Genetics, Section of Statistics, Microbiology, and Statistics, Faculty of Biology, University of Barcelona, Barcelona, Spain.*
[2] *Department of Clinical Foundations, School of Medicine and Health Sciences, University of Barcelona, Barcelona, Spain.*
[3] *Department of Administration and Political Science, Faculty of Law and Social Sciences, University of Talca, Chile.*

World Journal of Advanced Research and Reviews, 2023, 17(01), 531–536

Publication history: Received on 08 December 2022; revised on 16 January 2023; accepted on 19 January 2023

Article DOI: https://doi.org/10.30574/wjarr.2023.17.1.0098

Abstract

Background: The COVID-19 pandemic has had a global impact. Knowing the variables that affect the increase in infection is crucial for public health decision-making. Mobility and socio-demographic conditions of the population are important factors in the transmission of the SARS-CoV-2. The objective of this study is to analyze the relationship between people mobility, social determinants of health and COVID-19 cases using a Random Forest (RF) method.

Methods: The COVID-19 cases were analyzed in the Maule Region, Chile. Spearman rank was performed to analyze the total mobility index for each municipality. RF regression was used to create a model between COVID-19 infections, mobility index and sociodemographic variables. P-value <0.05 was considered statistically significant.

Results: Total mobility was highly correlated with new COVID-19 cases, adjusted for total population, in each municipality (ρ: 0.52-0.92). An upward trend is observed for mobility and COVID-19 cases for the 30 municipalities analyzed. For the RF model, COVID-19 active cases, total mobility, and external mobility are obtained as VIM. The most relevant demographic variables were overcrowding, density and area of municipality. The R-Squared was 0.68 for the performed RF model.

Conclusions: Artificial Intelligence methodologies are increasingly used for their excellent performance. RF Regression offers a clear solution for the design of predictor variables on the number of new cases per week. Mobility is a powerful predictor variable for the number of COVID-19 new cases.

Keywords: Random Forest; Mobility; Pandemics; COVID-19

1. Introduction

SARS-CoV-2 is a new type of coronavirus (a broad family of viruses that normally affect only animals) that can affect people and causes COVID-19. It was detected for the first time in December 2019 in Wuhan (China). COVID-19 has affected multiple countries on all continents. In addition, SARS-CoV-2 is highly infectious and can be transmitted through person-to-person contact and through direct contact with respiratory droplets generated when an infected person cough (1). The COVID-19 pandemic has highlighted how the disease is deepening in the most vulnerable populations with the greatest economic, educational, and social poverty. The social determinants of health are relevant to consider for managing the pandemic (2). Furthermore, during the first year of the pandemic it was critical to control

* Corresponding author: Nicolas Ayala 0000-0002-3530-6734

the mobility of the population to keep under control the number of new cases, hospitalizations, and patients in critical services (3).

Artificial Intelligence (AI) provides tools for predicting cases and classification systems in epidemiology and medical practice. The AI applications has been growing and multiple models have been developed for disease prediction considering population risk factors. Some of the most widely used models are Logistic Regression, Decision Trees (DT), and Random Forest (RF). An interesting approach is to use Artificial Intelligence methodologies to perform regression modelling to provide a comprehensive approach about the COVID-19 transmission. This provides information for make decisions in public health issues such as the use of masks, physical distancing, and lockdown of communities. Machine Learning methodologies is useful to study the relation between the covariables on the responding variable.

RF-regressors are regression tree models (Decision Trees), each trained with different subsets of input data, which reduces the variance of the predictors and minimizes overfitting. RF-regressors are highly effective at extracting non-linear relationships from input data while being time efficient. RF models have been used successfully in previous studies to predict foodborne disease, dengue, flu, and West Nile virus. For the last two years, RF has been widely used in models to forecast COVID-19 cases and to study variables of importance as risk factors (4).

We aimed to determine how COVID-19 dynamic variables are affected by population mobility and socio-demographic indicators using a RF model in the Maule region of Chile.

2. Material and methods

2.1. Area of Study

Chile is one of the most unequal countries in the world in terms of education and socio-economic income. Maule is a region located in Chile and it is one of the most vulnerable regions. In addition, Maule has a population of 1.044.950 population with 30 municipalities and represents 5,4% of Chile's population. Its area is 30.269 Km^2 and the density is 34,49 inhabitants/Km^2. Moreover, the main economic activities in the Maule region are agriculture and viticulture. A further concern aspect to consider is the highest poverty index in Maule region. Indeed, 225.728 inhabitants live in multidimensional poverty.

Weekly data, from each of the 30 municipalities from March to November 2020, were used for the statistical analysis. The current study analyses the data of COVID-19 rate cases (response variable), deaths by COVID-19, total deaths, and mobility. Mobility was measured by the GPS of people's mobile phones.

The mobility indices consider anonymized data for the period. Anonymized and aggregated telephone records were used to estimate the number of trips between municipalities. Based on this, it is important to note that this dataset does not give the exact location of the devices but rather the antenna to which it was connected. National and international privacy policies were respected.

Internal mobility index was measured, i.e., when people moved within the municipality. External mobility index was also measured, i.e., when people moved between their municipality of residence and another municipality. For the demographic and socioeconomic variables, we analyzed the total population (inhabitants), density (inhabitants/ Km^2), lack of basic services (%), income poverty (%), multidimensional poverty (%), and overcrowding population (%). Datasets of cases, mobility and sociodemographic variables were obtained from the Science Ministry of Chile as an open source (5).

3. Methodology of the analysis

3.1. Spearman Rank

The Spearman analysis was carried out to determine the relationship between mobility variables and COVID-19 infections for each municipality in the Maule region. When Spearman's rank coefficient ($\boldsymbol{\rho}$) tends to 1, it means a perfect association of rank. Conversely, when Spearman's rank coefficient tends to -1, it means a perfect negative association of rank. There is no association between variables when Spearman's rank coefficient tends to 0. A p-value <0.05 was considered to reject the null hypothesis and accept the alternative hypothesis.

3.2. Random Forest

The RF approach is based on multiple DT whose outcome is considered to classify a sample according to a majority voting strategy. The RF method takes the majority class provided by all the individual DTs. To improve the RF performance with respect to a single DT, it is necessary that the trees in the ensemble are diverse. On the one hand, to achieve this diversity, the training set to design each DT is created by applying bootstrap with replacement. RF is performed considering the CART algorithm as well as the Decision Tree. The RF model can be expressed in the following equation:

$$\hat{h}_{R_1} = \frac{1}{q}\sum_{l=1}^{q} \hat{h}\,(x, \Theta_l)$$

Let $(\hat{h}\,(.,\Theta_1), \ldots, \hat{h}\,(.,\Theta_q))$ be a collection of tree predictors, with $\Theta 1, \ldots, \Theta q$ q i.i. random variables independent of L_n. The random forest predictor $\hat{h}_{RF}$ is obtained by aggregating this collection of random trees. RF Regression was performed to study variables of importance and their relation. We obtained the number of trees, Variables of Importance (VIM) and R-Squared of the model performed. Data analysis was performed in R Studio, the packages used were "RandomForestSRC" to compute RF Regression and "ggRandomForests" for graphical analysis.

4. Results

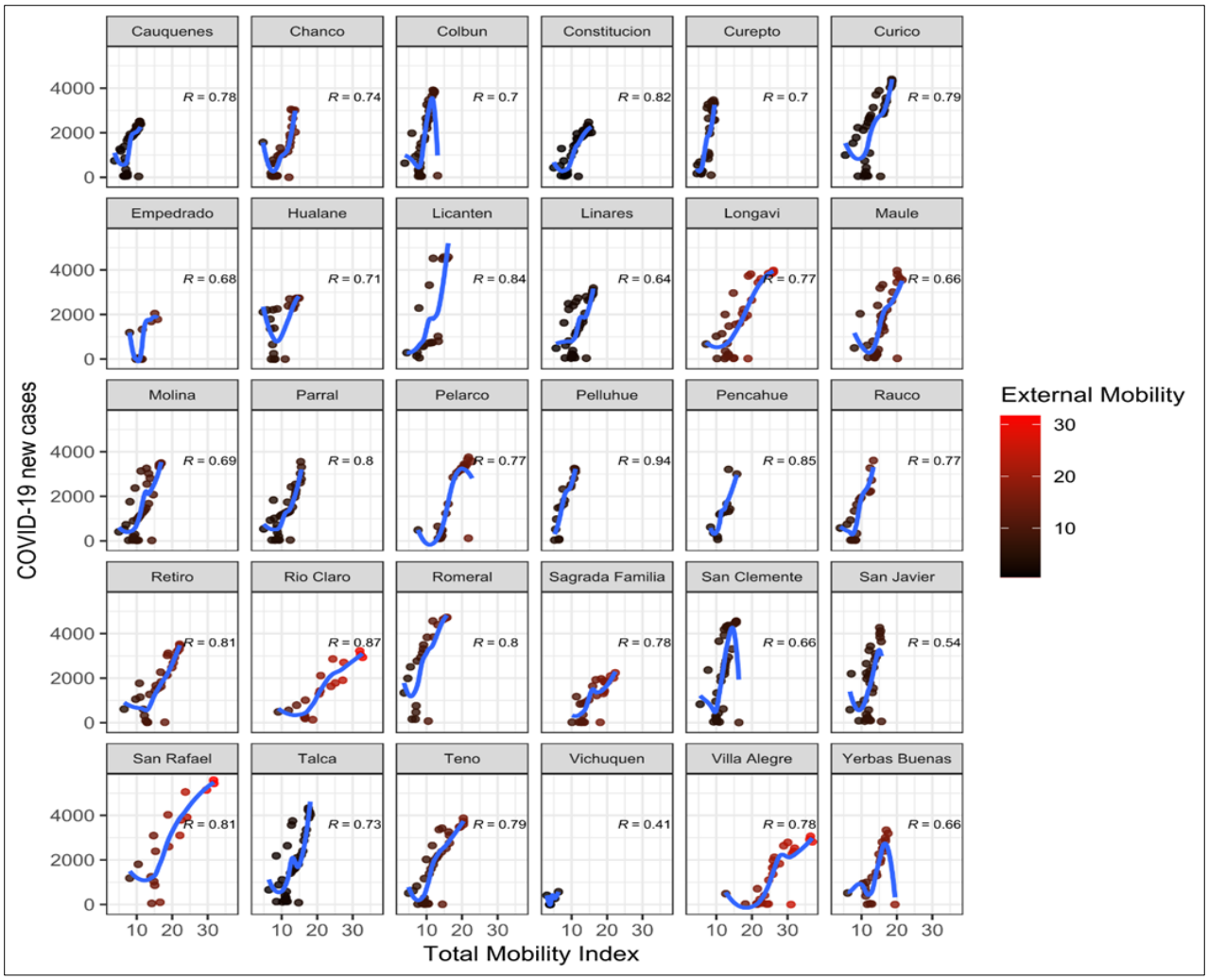


Figure 1 Total Mobility Index, External Mobility Index and COVID-19 new cases using LOESS analysis and Spearman Rank Correlation.

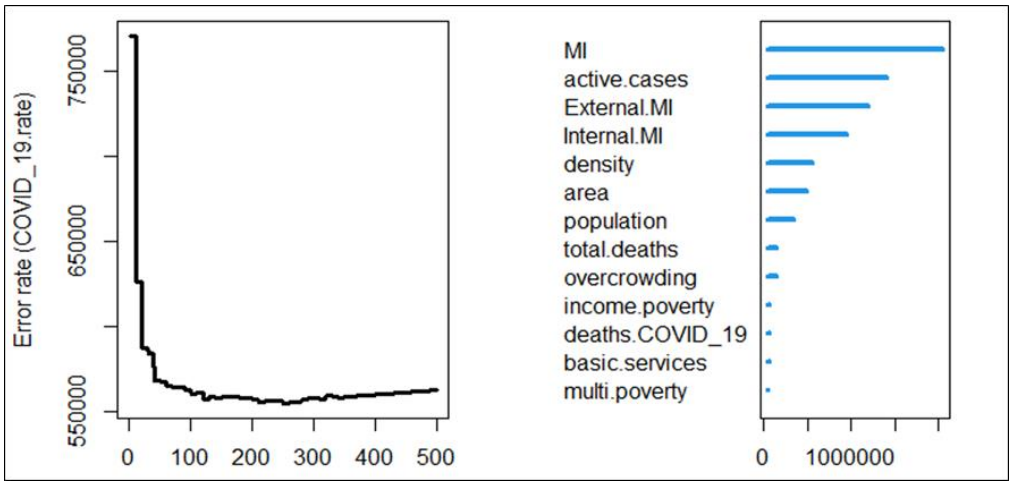
Figure 1 illustrates the evolution of COVID-19 cases and total mobility index in the Maule region by municipalities. There is a suggestive association between mobility and COVID-19 cases by municipality in Spearman Rank analysis. Spearman rank had a p-value <0.05 in all municipalities for the study of the relationship between total mobility index and COVID-19 case rate.

The range of Spearman rank results were (***ρ***: 0.41-0.94) for the cities of Vichuquén and Pelluhue, respectively. The highest Spearman correlation results were observed in the municipalities of Pelluhue (***ρ***: 0.92), Rio Claro (:0.87) and Pencahue (***ρ***: 0.85). On the other hand, the lowest Spearman ratios were observed in the municipalities of Vichuquén (***ρ***: 0.41), San Javier (***ρ***: 0.54) and Linares (***ρ***: 0.64). Municipalities with few banking and commercial services, close to cities with high-density populations, had a higher rate of external mobility; such as San Rafael, Villa Alegre, and Rio Claro. Highly populated cities such as Talca, Curicó, Linares and Cauquenes had the lowest external mobility index. The Spearman rank had a p-value <0.05 for all 30 municipalities.

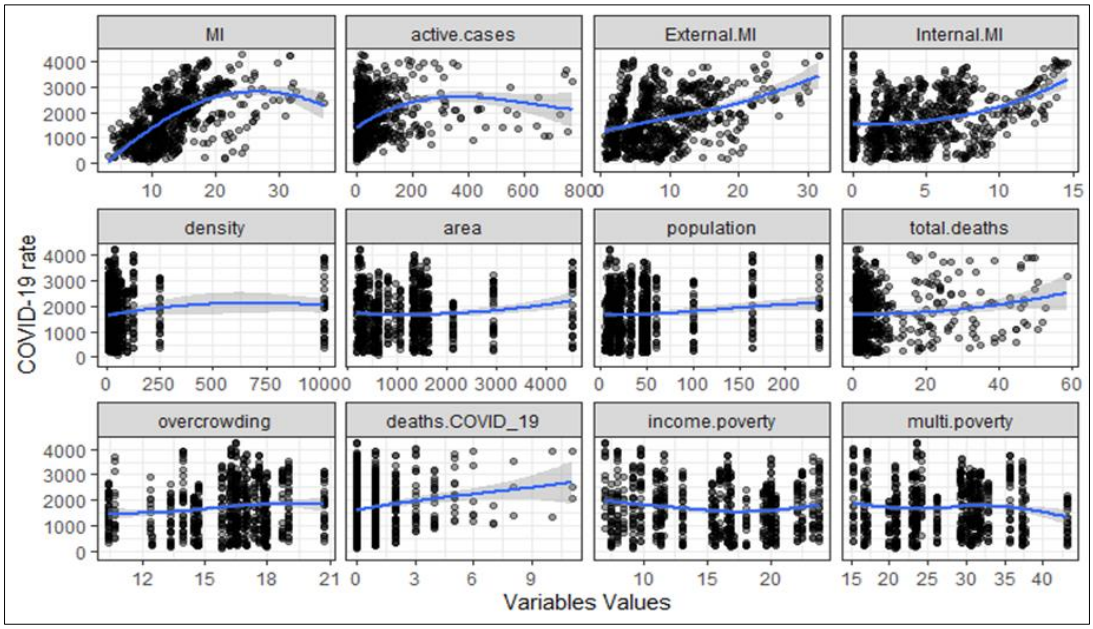
Random Forecast Regression was performed using 577 data in sample size. Number of variables tried at each split were 5, the average of terminal nodes was 73.65 and OBB R-Squared was 0.68. The Variable of Importance (VIM) for this model reveals that total mobility index, active cases, external mobility, and internal mobility are relevant to follow up during the pandemics. Regarding demographic variable, best performance was assigned for overcrowding population, density, surface, and income poverty. The regression model created with Random Forest sorted and prioritized the variables according to VIM.

Variable dependence plots illustrate the predicted response as a function of a covariate of interest, in this case, COVID-19 rate cases. Each observation is represented by a point on the plot and each predicted point is an individual observation, they dependent on the full combination of all other covariates such as mobility and sociodemographic topics. COVID-19 active cases show an upward curve that starts to stabilize after 200 active cases. In contrast, internal mobility and external mobility have a continuous upward behavior with respect to the caseload rate. Population overcrowding, municipality area and overcrowding have an ascending behavior as well as caseload but with a more moderate curve.

Random Forest model was used to estimate the regression of the COVID-19 case rate and the dynamic variables of the pandemic and demographics of the Maule region. **A.1** The lowest error rate (555939.1) is obtained with 180 trees. **A.2** Variable of Importance (VIM) for Random Forest Model. Most important variables are active total mobility index, active cases, extern mobility index and intern mobility index. **B.** Variable Dependence for COVID-19 rate cases in Maule Region. Graphs variables are sorted by VIM. Acronyms: Total Mobility Index (MI), External Mobility Index (External MI), Internal Mobility Index (Internal MI).



A.1 A.2



B

Figure 2 Random Forest Regression for COVID-19 rate cases in Maule Region.

5. Discussion

In the period of the data studied, the Ministry of Health applied different sanitary measures such as mandatory use of masks, physical distancing, and municipality lockdowns. In 2020, there were still not commercially available and approved vaccines to be used in the Chilean population. Monitoring mobility trends or corresponding variables could potentially inform mitigation measures towards slowing the spread of COVID-19. It can help to forecast the fast growth phase of COVID-19 pandemics considering demographic and socioeconomics variables for municipalities or regions. In the current research, RF was performed as a regression model to understand the relation of several variables such as mobility, COVID-19 deaths, active cases, socioeconomical factors and COVID-19 new cases.

Mobility and the dynamic of epidemic spread vary quite widely in many aspects. The growth and lag dynamics are different across different time scales such as days, weeks, or months. Mobility is highly correlated with the increase in the rate of COVID-19 infections per municipality. Hence the importance of quarantines in the early stage of the pandemic; total mobility, active cases, external mobility, and internal mobility are the most relevant variables in VIM in the RF model used. Similar situations were observed in the studies in the Americas, Europe, and Africa (6–8). Larger and more populated cities have a lower external mobility index when the index is adjusted for population size. The variables of demographic importance that performed best in the model were household overcrowding, land area, density, and income poverty.

Limitations

This study uses RF to explore a novel relationship between mobility, sociodemographic factors, and COVID-19 infection rates for the Maule region in Chile. The model studied was focused on a single place in Chile, the Maule Region. In addition, outliers could generate a high variance in the model. This is difficult to manage in such a long time series. Moreover, this study does not consider other factors like lockdown strategies, cases of concerned COVID-19 viruses and population age. There are several factors that could be considered in future models or projections.

The mobility data considers the distance that people move, but it leaves out the number of trips of the population. Finally, Socio-demographic factors could have been more clearly related in the model if the data were assigned to each case.

6. Conclusion

The COVID-19 pandemic has implied an additional effort in governments and healthcare centers to keep it under control. AI could be a useful tool because provide results with high performance. In this research case, Random Forest

methodology was used to generate a regression model and classify the variables with the greatest importance in the appearance of new cases of COVID-19 in each commune per week. We found high correlation between total mobility index and COVID-19 cases for each municipality of the Maule region in non-parametric analysis. Random Forest Model reveals that the variables of importance for the model were active cases, total mobility index, intern mobility index and extern mobility index.

Compliance with ethical standards

Acknowledgments

This work was supported by the University of Barcelona, Catalonia, Spain.

Disclosure of conflict of interest

All authors have no conflict of interest to declare.

References

[1] Sharma A, Ahmad Farouk I, Lal SK. COVID-19: A Review on the Novel Coronavirus Disease Evolution, Transmission, Detection, Control and Prevention. Viruses [Internet]. 2021 Feb 1 [cited 2022 Oct 4];13(2). Available from: https://pubmed.ncbi.nlm.nih.gov/33572857/

[2] Morante-García W, Zapata-Boluda RM, García-González J, Campuzano-Cuadrado P, Calvillo C, Alarcón-Rodríguez R. Influence of Social Determinants of Health on COVID-19 Infection in Socially Vulnerable Groups. Int J Environ Res Public Health [Internet]. 2022 Feb 1 [cited 2022 Oct 4];19(3). Available from: https://pubmed.ncbi.nlm.nih.gov/35162317/

[3] Nouvellet P, Bhatia S, Cori A, Ainslie KEC, Baguelin M, Bhatt S, et al. Reduction in mobility and COVID-19 transmission. Nature Communications 2021 12:1 [Internet]. 2021 Feb 17 [cited 2022 Oct 4];12(1):1–9. Available from: https://www.nature.com/articles/s41467-021-21358-2

[4] Alballa N, Al-Turaiki I. Machine learning approaches in COVID-19 diagnosis, mortality, and severity risk prediction: A review. Inform Med Unlocked [Internet]. 2021 Jan 1 [cited 2022 Oct 4];24. Available from: https://pubmed.ncbi.nlm.nih.gov/33842685/

[5] MinCiencia/Datos-COVID19: Para señalar fuente de los datos señalar que vienen de este repositorio, junto con la fuente de orígen: "Datos obtenidos desde el Ministerio de Ciencia y producidos por el Ministerio de Salud (o la fuente que corresponda) https://github.com/MinCiencia/Datos-COVID19". Please attribute data provenance: produced by Chile Ministry of Health and obtained from Ministry of Science https://github.com/MinCiencia/Datos-COVID19" [Internet]. [cited 2022 Oct 1]. Available from: https://github.com/MinCiencia/Datos-COVID19

[6] Zhu G, Stewart K, Niemeier D, Fan J, Manley E, Delmelle E, et al. Understanding the Drivers of Mobility during the COVID-19 Pandemic in Florida, USA Using a Machine Learning Approach. ISPRS International Journal of Geo-Information 2021, Vol 10, Page 440 [Internet]. 2021 Jun 28 [cited 2022 Oct 4];10(7):440. Available from: https://www.mdpi.com/2220-9964/10/7/440/htm

[7] Araújo F, Araújo F, Machado K, Rosário D, Cerqueira E, Villas LA. Ensemble mobility predictor based on random forest and Markovian property using LBSN data. Journal of Internet Services and Applications [Internet]. 2020 Dec 1 [cited 2022 Oct 4];11(1):1–11. Available from: https://jisajournal.springeropen.com/articles/10.1186/s13174-020-00130-7

[8] De Palma A, Vosough S, Liao F. An overview of effects of COVID-19 on mobility and lifestyle: 18 months since the outbreak. Transp Res Part A Policy Pract. 2022 May 1;159:372–97.