GOVERNMENT MEASURES AGAINST THE COVID-19 PANDEMIC MUST BE DETERMINED ACCORDING TO THE SOCIO-ECONOMIC STATUS OF THE COUNTRY

Kathleen Lois Foster [*]
Department of Biology
Ball State University
Muncie, IN 47306, USA
klfoster@bsu.edu

Alessandro Maria Selvitella
Department of Mathematical Sciences
Purdue University Fort Wayne
Fort Wayne, IN, 46305
aselvite@pfw.edu

ABSTRACT

This study uses machine learning methods to investigate the relationship between socio-economic determinants pre-dating the pandemic and the reported number of cases, deaths, and the ratio of deaths/cases in 199 countries/regions during the first months of the COVID-19 pandemic. We generate a portfolio/ensemble of 32 interpretable models and consider both the case in which the epidemiological variables are independent and the case in which their dependence is weighted based on geographical proximity. We build a measure of variable importance, the *Signed Importance Index* (SII) whose role is to identify the most contributing socio-economic factors to the variability of the COVID-19 pandemic. Our results suggest that, together with the established influence on cases and deaths of the level of mobility, the specific features of the health care system (smart/poor allocation of resources), the economy (equity/non-equity), and the society (religious/not religious or community-based vs not) contributed to the number of COVID-19 cases and deaths heterogeneously across countries early in the pandemic.

1 INTRODUCTION

Since the outbreak of the COVID-19 pandemic (World Health Organization), several studies (Hilton & Keeling, 2020; Tuite et al., 2020; Giordano et al., 2020) have concentrated on the biological and epidemiological factors governing COVID-19 transmission, while few others (Stojkoski et al., 2020; Selvitella & Foster, 2020) have investigated the potential impact of socio-economic characteristics on governing the extent of COVID-19 diffusion in the population. Societal and economic factors can be of critical importance for accuracy of models of the outbreak because of the economic (Mikhael & Al-Jumaili, 2020; Di Marco et al., 2020) and health impacts of the drastic measures that have been put in place in an effort to slow the spread of the disease (e.g social distancing, quarantine, lockdowns, testing, and reallocation of hospital resources) (Ferguson et al., 2020; Tian et al., 2020).

In this paper, we take a reverse perspective and analyze how socio-economic determinants pre-dating the pandemic relate to the number of reported cases, deaths, and the ratio deaths/cases of COVID-19 via machine learning methods. Our focus is on understanding the connection between epidemiological variables of the COVID-19 pandemic and the (i) Capacity of a country to deal with COVID-19 cases (Healthcare Infrastructure); (ii) Statistics indicative of the health of the population (Health Statistics); (iii) Economic situation and tourism/mobility (Economic Health); (iv) Demographic structure, in particular the age structure and the spatial distribution of the population (Demographic Structure); (v) Societal characteristics such as the level of education, access to technology, and features of government (Societal Characteristics); (vi) Pollution level and ecological footprint (Environmental Health); and (vii) Religious practices (Religious Characteristics). We hypothesize that different countries have different specific socio-economic features and therefore the efficacy of government measures and the incidence of the disease must be smart and heterogeneous across countries and across resources.

[*] Both authors contributed equally in this work.

We analyze 32 interpretable models, including (i) regression models with both independent and proximity dependent outcomes; and (ii) variable selection through LASSO. Then, we build a Signed Importance Index (SII) to help focus only on those findings that are common to a majority, thereby reducing the sensitivity of our findings to the limitations of any single model considered.

2 Materials and Methods

2.1 Outcome variables Y: Epidemiological Variables

The total number of reported cases and deaths attributed to COVID-19 as of 2nd May 2020 were obtained from Our World In Data (Our World In Data) and used as outcome variables of our models.

2.2 Predictors X: Socio-economic (SE) factors

Forty-four SE variables were chosen for our analyses based on their potential explanatory power and to facilitate comparisons with other published works (Stojkoski et al., 2020; Selvitella & Foster, 2020). Data were obtained from publicly available databases (World Bank Open Data; International Monetary Fund; United Nations; Global Footprint Network; State of Global Air) for a total of 199 countries/regions, 32 of which only had data for all 44 variables of interest. As the years for which data were available varied by country/region, we chose to use the most recent data available for each country, (oldest being 2010, most recent being 2019).

2.3 Weighting Matrix A: Geographic Information

Latitude and longitude coordinates for capital cities were obtained for each country from the CEPII (Centre d'Études Prospectives et d'Informations Internationales) GeoDist database. Coordinates for 11 countries were not found in the GeoDist database and were obtained from Google.

Using these coordinates, we calculated the pairwise distances between cities using the *Spherical Law of Cosines*. Given two cities C_1, C_2 on the surface of the Earth with latitude and longitude coordinates (α_1, β_1) and (α_2, β_2), and assuming a constant radius of the Earth $R = 6378$ kms, we have that the distance between C_1 and C_2 is given by the following formula

$$d(C_1, C_2) = arc\cos(\sin(\alpha_1) * \sin(\alpha_2) + \cos(\alpha_1) * \cos(\alpha_2) * \cos(\beta_2 - \beta_1)) * R.$$

The matrix A has then been computed such that the entries were given by

$$A_{ij} = \frac{\exp\{-d(C_i, C_j)\}}{\exp\{-d(\cdot, \cdot)\}},$$

for $i, j = 1, \cdots, 199$. Here $d(\cdot, \cdot)$ is a normalization factor computed as the average over the distances between every city C_i and every city C_j. We considered the ellipticity of the Earth to have a minor influence on our results and we considered the spherical approximation appropriate. Our strategy relates to theoretical work (Daskalakis et al., 2019) on regression with correlated responses.

2.4 MICE

Missing SE determinant data were imputed using the R package *mice*, which performs imputation via *Multivariate Imputation by Chained Equations* (MICE). This method, fully described in (Van Buuren & Groothuis-Oudshoorn, 2011; Azur et al., 2011), assumes that the probability that a value is missing depends only on observed values and not on unobserved values. We assumed this throughout all our analyses.

2.5 Regression Models

We considered 5 different outcome variables: (i) Y_1= # cases; (ii) Y_2= # deaths; (iii) $\tilde{Y}_1$= # cases/total population; (iv) $\tilde{Y}_2$ = # deaths/total population; and (v) Y_0 = # deaths/# cases. We considered two sets of explanatory variables: (i) All variables ($|\mathbf{X}| = 44$); and (ii) All but total population (POP) count ($|\mathbf{X}| = 43$). Models for automatic variable selection, such as LASSO (Tibshirani, 1996; James et al., 2013) were also considered. We used R Studio Version 1.2.5042 and libraries *readxl*, *readr*, *gdata*, *mice*, *glment*, *caret* for the computations.

2.6 Importance index SII

To measure the importance of the variables across our models, we built a *Signed Importance Index* (SII), which counts the presence of a variable in the top-10 correlated variables with sign. For example if the variable X appears 12 times in the top-10 highest correlated variables with Y, 10 times positively correlated, then $SII = 8$. We computed this index globally and for each single type of outcome variable.

3 Results and Discussion

3.1 Healthcare Infrastructure and Statistics

The number of physicians, essential health coverage index, and death rate were among the top-10 variables in 25%, 25%, 15.63% of our models, respectively (Figure 1A,B). Physicians correlated positively with all models except Y_0. Access to essential health services and crude death rate correlated positively in our models, appearing in $25 - 50\%$ of Y_1, Y_2, and Y_0 models. Number of nurses and midwives, number of hospital beds, prevalence of diabetes, and crude birth rate correlated negatively in 37.5%, 50%, 21.88%, and 43.75% of our models, respectively (Figure 1A,B). Hospital beds and diabetes appeared in $25 - 75\%$ of all model categories, with diabetes absent from Y_1 models. Nurses/midwives were important in 100% of $\tilde{Y}_1$ models and 50% of $\tilde{Y}_2$ models, exclusively. Birth rate was important in 25% of Y_1 and $\tilde{Y}_1$ models and 50% of $\tilde{Y}_2$ models. Although seemingly contradictory, taken together, these data may provide indications that government healthcare spending needs to be allocated heterogeneously in order to effectively combat diseases like COVID-19.

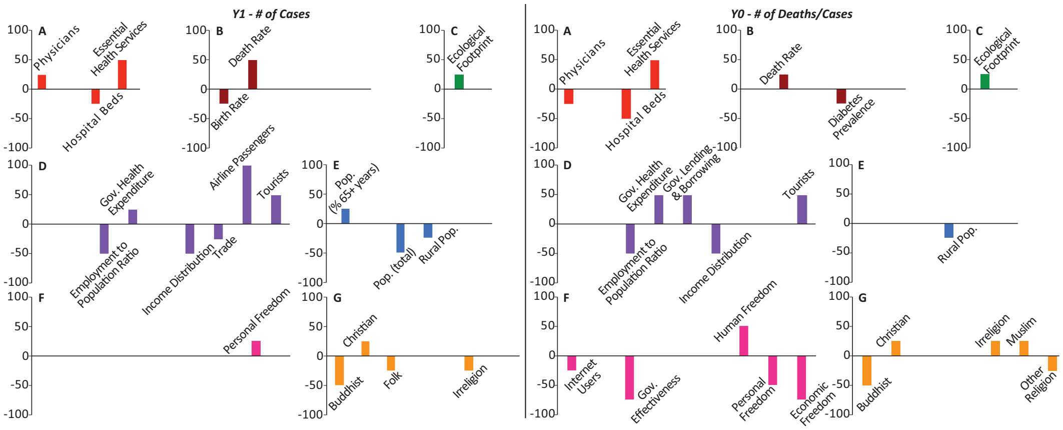

Figure 1: Indices of importance SII of socio-economic variables in models of number of COVID-19 cases (Y_1; LEFT), and deaths/cases (Y_0; RIGHT). We analyzed the effects of 44 socio-economic determinants in a total of 4 (Y_1) and 8 (Y_0) models. The number of times each variable appeared among the top-10 highest correlated variables in these models was tallied with its own sign (sign + if positively correlated, sign - if negatively correlated) and the signed percentage of this tally is shown by the bars above. A: Health Infrastructure, B: Health Statistics, C: Environmental Health, D: Economic Health, E: Demographic Structure, F: Societal Characteristics, and G: Religion.

3.2 Economic Health

As shown in Figure 1D, domestic government health expenditure was the only measure of economic health that was identified as important and correlated positively in all models. In contrast, employment to population ratio and income distribution consistently correlated negatively with the outcome variables in all models and were identified as important determinants in 43.75% of all our models. This suggests that countries with higher levels of employment and less economic disparity, can be expected to feel the effects of COVID-19 less strongly. GDP, trade, and government lending/borrowing correlated inconsistently across the models, with GDP and trade only identified as important variables in models lacking the geographical weighting matrix.

3.3 Demographic Structure and Mobility

Variables relating to demographic structure consistently played a smaller role in Y_0 than the other models (Figure 1E). Total population correlated negatively with Y_1 and Y_2. Population density, the proportion of the population over the age of 65, and the proportion of immigrants correlated positively and were important factors in 28.13%, 25%, and 31.25% of our models, though never in Y_0. Rural population correlated negatively in Y_1, Y_2, and Y_0 models and positively in $\tilde{Y}_1$ and $\tilde{Y}_2$ models. The number of airline passengers per year was nearly always important for Y_1 and Y_2 models, correlating positively in 100% and 75% of those models, respectively (Figure 1D). Although the number of tourist arrivals was consistently positively correlated with 50% of Y_1, Y_2, and Y_0 models, the direction of correlation was inconsistent in $\tilde{Y}_1$ and $\tilde{Y}_2$, despite being identified as important in 75% and 100% of those models, respectively (Figure 1D). The geographical weighting matrix was always present in models in which tourist arrivals correlated negatively, while the absence of this matrix in the models always coincided with tourist arrivals correlating positively.

3.4 Environmental Health

Ecological footprint was an important variable in 25% of Y_1, Y_2, and Y_0 models, in which it consistently correlated positively (Figure 1C). Air pollution never appeared as an important determinant.

3.5 Societal and Religious Characteristics

The majority of the societal variables that correlated strongly in our models had a net negative correlation (Figure 1F). Access to internet was identified as an important determinant in 28.13% of all our models; in $\tilde{Y}_1$, $\tilde{Y}_2$, and Y_0 models, it correlated negatively, whereas in Y_2 models it correlated positively. Government effectiveness and economic freedom score correlated negatively and were important in 75% of Y_0 models. Personal freedom also correlated negatively with Y_0 in 50% of those models, though it correlated positively in one Y_1 model. Human freedom correlated positively in 50% of Y_0 models, but negatively in 25% of $\tilde{Y}_1$ and $\tilde{Y}_2$ models. The average number of people per household appeared in 25% of $\tilde{Y}_1$ models, in which it correlated positively. The percentage of the population identifying as Christian was a positive correlate and appeared in 40.63% of models (Figure 1G). In contrast, the percentage of the population identifying as Buddhist correlated negatively and appeared in 50% of Y_1, Y_2, and Y_0 models. The remaining religious categories appeared rarely among the most important variables in our models, but when present, most correlated negatively.

4 Conclusions

Although our analyses do not yet incorporate such complex factors as the time evolution of the disease or underreporting (Li et al., 2020; Hilton & Keeling, 2020; Jagodnik et al., 2020; Shutta et al., 2017), our results suggest that governments might need to allocate healthcare resources heterogeneously, with a possible benefit in decentralizing healthcare. This could be a problem for developing countries, where the means are limited. As of May 2nd, 2020, countries with more economic equity among their citizens seemed less hit by COVID-19, possibly indicating the importance of having a minimal baseline assistance across the whole population of a country. The reduced degree of mobility across countries, for example the degree to which tourism is constrained, had a positive effect in reducing the number of cases, deaths, and death rate per cases. However, there is an indication that a smart and alternating policy could lead to further containment of the disease. Furthermore, our analysis highlighted the benefit of informing the population for government measures to be more effective. Together, our results seem to indicate that blanket policies are sub-optimal and government measures related to healthcare and immigration have the potential to both help and damage the population, as, if not appropriately taken, they can lead to an increase or reduced decrease of COVID-19 cases, deaths, and deaths/cases rate.

Acknowledgments

KLF and AMS would like to thank their families for their constant support. This work was funded through start-up grants from Ball State University and Purdue University Fort Wayne.

REFERENCES

M. J. Azur, E. A. Stuart, C. Frangakis, and J. L. Leaf. Multiple imputation by chained equations: what is it and how does it work? *International Journal of Methods in Psychiatric Research*, 20: 40–49, 2011.

C. Daskalakis, N. Dikkala, and I. Panageas. Regression from dependent observations. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing*, STOC 2019, pp. 881–889, New York, NY, USA, 2019. Association for Computing Machinery.

M. Di Marco, M. L. Baker, P. Daszak, P. D. Barro, E. A. Eskew, C. M. Godde, T. D. Harwood, M. Herrero, A. J. Hoskins, E. Johnson, W. B. Karesh, C. Machalaba, J. N. Garcia, D. Paini, R. Pirzl, M. S. Smith, C. Zambrana-Torrelio, and S. Ferrier. Opinion: Sustainable development must account for pandemic risk. *Proceedings of the National Academy of Sciences*, 117:3888––3892, 2020.

N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Cucunubá Z., G. Cuomo-Dannenburg, A. Dighe, I. Dorigatti, H. Fu, K. Gaythorpe, W. Green, A. Hamlet, W. Hinsley, L. C. Okell, S. van Elsland, H. Thompson, R. Verity, E. Volz, H. Wang, Y. Wang, P. G. Walker, C. Walters, P. Winskill, C. Whittaker, C. A. Donnelly, S. Riley, and A. C. Ghani. Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand. *Imperial College COVID-19 Response Team*, pp. 1–20, 2020.

G. Giordano, F. Blanchini, R. Bruno, P. Colaneri, A. D. Filippo, A. D. Matteo, and M. Colaneri. Modelling the covid-19 epidemic and implementation of population-wide interventions in italy. *Nature Medicine*, 26:855–860, 2020.

Global Footprint Network. URL `http://data.footprintnetwork.org/#/`.

J. Hilton and M. J. Keeling. Estimation of country-level basic reproductive ratios for novel coronavirus (covid-19) using synthetic contact matrices. *medRxiv*, pp. 1–7, 2020.

International Monetary Fund. URL `https://www.imf.org/en/data`.

K. M. Jagodnik, F. Ray, F. M. Giorgi, and A. Lachmann. Correcting under-reported covid-19 case numbers: estimating the true scale of the pandemic. *medRxiv*, pp. 1–6, 2020.

G. James, D. Witten, T. Tastie, and R. Tibshirani. *Introduction to Statistical Learning*. Springer, 2013.

R. Li, S. Pei, B. Chen, Y. Song, T. Zhang, W. Yang, and J. Shaman. Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, 368:489–493, 2020.

E. M. Mikhael and A. A. Al-Jumaili. Can developing countries alone face coronavirus? an iraqi situation. *Public Health in Practice*, 1:100004, 2020.

Our World In Data. URL `http://ourworldindata.org/`.

A. M. Selvitella and K. L. Foster. Societal and economic factors associated with covid-19 indicate that developing countries could suffer the most. *Technium Social Sciences Journal*, 10:637–644, 2020.

D. P. Shutta, C. A. Manorec, S. Pankavich, A. T. Porter, and S. Y. D. Valle. Estimating the reproductive number, total outbreak size, and reporting rates for zika epidemics in south and central america. *Journal of the Royal Society Interface*, 21:63—-79, 2017.

State of Global Air. URL `https://www.stateofglobalair.org/engage`.

V. Stojkoski, Z. Utkovski, P. Jolakoski, D. Tevdovski, and L. Kocarev. The socio-economic determinants of the coronavirus disease (covid-19) pandemic. *medRxiv*, pp. 1–22, 2020.

H. Tian, Y. Liu, Y. Li, C.-H. Wu, B. Chen, M. U. G. Kraemer, B. Li, J. Cai, B. Xu, Q. Yang, B. Wang, P. Yang, Y. Cui, Y. Song, P. Zheng, Q. Wang, O. N. Bjornstad, R. Yang, B. T. Grenfell, O. G. Pybus, and C. Dye. An investigation of transmission control measures during the first 50 days of the covid-19 epidemic in china. *Science*, 368:638–642, 2020.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, 58:267–288, 1996.

A. Tuite, D. N. Fisman, and A. L. Greer. Mathematical modeling of covid-19 transmission and mitigation strategies in the population of ontario, canada. *medRxiv*, pp. 1–23, 2020.

United Nations. URL `https://population.un.org/Household/index.html#/countries/840`.

S. Van Buuren and K. Groothuis-Oudshoorn. mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, 43:1–67, 2011.

World Bank Open Data. URL `https://data.worldbank.org/`.

World Health Organization. Rolling updates on coronavirus disease (covid-19). URL `https://www.who.int/emergencies/diseases/novel-coronavirus-2019/events-as-they-happen`.