TECHNIUM
SOCIAL SCIENCES JOURNAL

9R O

$ QHZ GHFD
IRU VRFLDO

, 66 1

Z Z Z W H F K Q L X P V F L H

TECHNIUM SOCIAL SCIENCES JOURNAL

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

Societal and economic factors associated with COVID-19 indicate that developing countries could suffer the most

Alessandro Maria Selvitella

Department of Mathematical Sciences
Purdue University Fort Wayne
aselvite@pfw.edu

Kathleen Lois Foster

Department of Biology
Ball State University
klfoster@bsu.edu

Abstract. Most of the research related to the COVID-19 pandemic deals with the biological and epidemiological factors which have driven the spread of the coronavirus around the globe. In this paper, we analyse how societal and economic variates relate to the number of cases and deaths across countries, via machine learning methods. Our findings recommend focusing our attention on developing countries where the healthcare system might suffer the most.

Keywords. COVID-19, socio-economic determinants, explainability, variable selection, LASSO.

1. Introduction

Since the beginning of 2020, the novel coronavirus (SARS-CoV-2) has hit basically the whole globe and dramatically impacted our daily lives. The number of cases and deaths caused by the virus, associated comorbidities, and their socio-economic consequences have captured the attention of every country in the world, starting from the most developed, such as China, USA, and all European countries.

Officially, the first case of the novel coronavirus was reported in China on 31 December 2019. Since then it spread globally and it is still among us at the time of writing of this manuscript. The WHO publicly confirmed the coronavirus a Public Health Emergency of International Concern on 30 January 2020 and later on a pandemic on 11 March 2020 [25].

The attention of researchers has been focused on the biological and epidemiological factors governing COVID-19 transmission [8,21,6] with little investigation into relevant socio-economic factors (but see [18]). We need to realize that these socio-economic factors can help better characterize the evolution of the pandemic [14,4], as the power of response to the pandemic with which each country can strike back against the virus is strongly related to the societal characteristics and to the economy of the country. The health measures put in place by each government in order to fight the pandemic, such as social distancing, quarantine,

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

lockdowns, etc. [5,19] might have different effects depending on the level of development of the country, its economy, and the type of society in which they are adopted.

This work is dedicated to understanding how the relationship between the number of cases and deaths and the socio-economic characteristics of a country can provide a better understanding on the evolution of the disease in developing countries where the healthcare system is possibly less prepared to defend the population during the time of the COVID-19 pandemic. Our findings support social distancing measures and travel restriction, and determine that efforts are needed to protect countries where the health-care system might not be ready to deal with the pandemic.

The remaining part of the paper is organized as follows. In Section 2, we discuss the features of the data that we analysed (Table 1); in Section 3 we provide the details of our statistical analysis; Section 4 is dedicated to our results and discussion. Section 5 is dedicated to our models limitations and Section 6 is dedicated to the conclusions of our analysis.

2. Datasets

In this section, we collect information about the variables included in our statistical data analysis and some of the features of our datasets. Please, refer to Table 1 for more information about the sources of the data analysed.

We considered four different outcome variables for our analysis: Y_1= # cases, Y_2= # deaths, $\tilde{Y}_1$= # cases/total population and $\tilde{Y}_2$ = # deaths/total population. The total number of reported cases and deaths attributed to COVID-19 as of 2nd May 2020 were obtained from Our World In Data [15] to use as outcome variables of our models.

Explanatory variables were obtained from publicly available databases [24,9,22,7,17] which included a total of 175 countries/regions, 59 of which had data for all 30 variables of interest. We analyzed two sets of explanatory variables: |X| = 30, namely including all all variables and |X| = 29, where the total population count (POP) was removed. The SE variables were chosen for our analyses based on their potential explanatory power and to facilitate comparisons with other published work [18]. As the years for which data were available varied by country/region, we chose to use the most recent data available, between 2010 and 2019.

SE factors were divided into six categories according to 1) the capacity to deal with COVID-19 cases (Healthcare Infrastructure), 2) statistics indicative of the health of the population (Health Statistics), 3) economic situation, tourism/mobility, and policy (Economic Health/Mobility), 4) societal characteristics related to the education, access to technology, government (Societal Characteristics), 5) basic demography related to age structure and spatial distribution of the population (Demographic Structure), and 6) pollution/ecological footprint (Environmental Health). Please, refer to Table 1 for more details.

Although our complete dataset contained 175 countries/regions, missing values resulted in preliminary linear regression [12] models excluding 116 of those countries/regions. We imputed the missing values using MICE [23,1].

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

Table 1. Data sources for socio-economic determinants.

Variable	*Source*	*Link*
Healthcare Infrastructure		
Physicians (/1000 people)	World Bank Open Data	https://data.worldbank.org
Nurse and midwives (/1000 people)	World Bank Open Data	https://data.worldbank.org
Hospital beds (/1000 people)	World Bank Open Data	https://data.worldbank.org
Essential health services (UHC) coverage index	World Bank Open Data	https://data.worldbank.org
Health Statistics		
Birth rate (crude, /1000 people)	World Bank Open Data	https://data.worldbank.org/
Death rate (crude, /1000 people)	World Bank Open Data	https://data.worldbank.org/
Life expectancy at birth (years)	World Bank Open Data	https://data.worldbank.org/
Prevalence of diabetes between ages 20-79 (% population)	World Bank Open Data	https://data.worldbank.org/
Mortality from unsafe water, or sanitation, lack of hygiene combined (/100k people)	World Bank Open Data	https://data.worldbank.org/
Completeness of death registration with cause-of-death information (%)	World Bank Open Data	https://data.worldbank.org/
Economic Health		
GDP (per capita, PPP, $)	World Bank Open Data	https://data.worldbank.org/
Unemployment rate (most recent available, % labor force)	International Monetary Fund	https://www.imf.org/en/data
Employment to population ratio for ages 15+ (modeled ILO estimate)	World Bank Open Data	https://data.worldbank.org/
Domestic general government health expenditure (per capita, PPP, $)	World Bank Open Data	https://data.worldbank.org/
Government lending/borrowing (% GDP)	International Monetary Fund	https://www.imf.org/en/data
Income distribution (GINI index)	World Bank Open Data	https://data.worldbank.org/
Trade (% GDP)	World Bank Open Data	https://data.worldbank.org/

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

Number of airline passengers (per year)	World Bank Open Data	https://data.worldbank.org/
Number of tourist arrivals (per year)	World Bank Open Data	https://data.worldbank.org/
Demographic Structure		
Population aged 65+ (% population)	World Bank Open Data	https://data.worldbank.org/
Population aged 0 − 14 (% population)	World Bank Open Data	https://data.worldbank.org/
Population (total)	World Bank Open Data	https://data.worldbank.org/
Rural population (% population)	World Bank Open Data	https://data.worldbank.org/
International migrant stock (% population)	World Bank Open Data	https://data.worldbank.org/
Population density (people per sq km)	World Bank Open Data	https://data.worldbank.org/
Environmental Health		
Ecological footprint (gha/person)	Global Footprint Network	http://data.footprintnetwork.org/#/
Air pollution (avg P.M. 2.5 exposure per year)	State of Global Air	https://www.stateofglobalair.org/ engage
Societal Characteristics		
Individuals using internet (% population)	World Bank Open Data	https://data.worldbank.org/
Education level: Human capital index (0 − 1)	World Bank Open Data	https://data.worldbank.org/
Avg number of persons per household	United Nations	https://population.un.org/Household/index.html#/countries/840

3. Explainable Models and Variable Selection Methods

We first analyzed the multivariate correlation of the explanatory variables through multivariate linear regression, whose results are reported below in Subsections 4.1 to 4.4.

Then, in order to understand the relative importance of our input variables we model the relationship between our explanatory variables X and the four outcome variables using LASSO. LASSO is a variable selection method based on an algorithm which minimizes the least-square-error of classical linear regression under a budget constraint on the L^1-norm of the coefficients.

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

Given the sample of n = 175 countries, and outcome variables y_i (one of our four outcome variables) each including p = 30 or p = 29 SE covariates $x_i = (x_1, \ldots, x_p)^T$, then LASSO optimizes the following objective function:

$$\min_{\beta_0,\beta} \left\{ \sum_{i=1}^{n} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t.$$

Here *t* is a budget parameter that determines the amount of regularization and the number of variables selected in the model. The reduced amount of variables provided by the LASSO with an appropriately small value of *t* outputs the socio-economic variable better associated with the number of reported cases and deaths due to COVID-19.

We used R Studio Version 1.2.5042 libraries *readxl, readr, gdata, mice, glment, caret* for our statistical analysis.

4. Results and Discussion

Our models with 59 countries (without imputation) were consistent with the results obtained from the 175-country imputed dataset. Below are the results of our multivariate analysis.

4.1. Demographic Structure.

POP correlates negatively with Y_1 and Y_2, but becomes not significant for $\tilde{Y}_1$ and $\tilde{Y}_2$, possibly suggesting a sub-exponential growth of the number of infections [2]. Life expectancy correlates positively with $\tilde{Y}_1$ and $\tilde{Y}_2$, but interestingly % POP +65 correlates negatively. POP density correlates positively with Y_2 and weakly with Y_1 as we would expect, but interestingly average household size correlates negatively (weakly) with Y_1 , speculatively as an indication of family care.

4.2. Environmental Health.

Level of pollution correlates weakly negatively with $\tilde{Y}_2$. This fact needs to be taken into account in compartmental models [13,3], especially as it is currently not understood if COVID-19 can be air-transmitted, or vector-borne. This result may indicate an influence of air transmissibility or air quality on COVID-19.

4.3. Economic Health/Mobility.

The % of POP composed of immigrants and refugees correlates negatively with Y_2 . Although the possible reasons for this are unknown, it may indicate a greater strength or resilience of a population with greater genetic diversity. Regardless of the reason, these results strongly suggest that greater immigrant populations are not a drain on health resources related to COVID-19 and thus contrasts with "fear of immigrants" policies that may argue for long-term border closure policies. In contrast, the average number of tourist arrivals and airline passengers per year positively correlate with the number of COVID-19 cases and deaths and support short-term travel restrictions. Interestingly, in contrast with [18], GDP did not correlate with COVID-19 cases and deaths.

4.4. Healthcare Infrastructure and Statistics.

Both Y_2 and $\tilde{Y}_2$ are impacted by SE factors related to social services (negatively correlated with number of available hospital beds). The number of deaths thus appears to be influenced by the availability of healthcare resources, confirming the importance of efforts to reduce healthcare overload and flattening the curve. This calls for particular care for developing countries [14], which can be hit heavily and can be a possible source of an uncontrolled second wave. Prevalence of diabetes correlates negatively with Y_2, when taking into

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

consideration all the variates, highlighting the possibility that COVID-19 has less impact in systems with existing infrastructure to provide access to care for chronic conditions. Singularly, diabetes is positively correlated with total cases and deaths, but negatively correlated with the number of beds and so its effect is possibly confounded by the inclusion of the other factors in our multivariate models.

4.5. Variable Selection.

LASSO selected trade (% GDP), government health expenditure (per capita), and government lending/borrowing (% GDP) in the top 10 variables of all models. Prevalence of diabetes was important for all but Y_1 and the number of hospital beds was important for all but $\tilde{Y}_1$. Pollution, short-term mobility (airline passengers and tourist arrivals), death rate, and unemployment rate were important for both Y_1 and Y_2, but disappeared from the same models when divided by population ($\tilde{Y}_1$, $\tilde{Y}_2$). This apparent discrepancy with our linear regression models may be due to strong correlations between these two variables and POP, and LASSO not selecting highly correlated variables at the same time. In contrast, the number of physicians, life expectancy at birth, and immigrant/refugee population (% POP) were identified in the top 10 variables only when the total number of cases and deaths were divided by POP.

Please find in Figure 1 above the LASSO-plot of the most important variables selected in the model with outcome variable Y_2 . Similar plots can be produced for the other models, by simple modifications of the R-codes (available upon request).

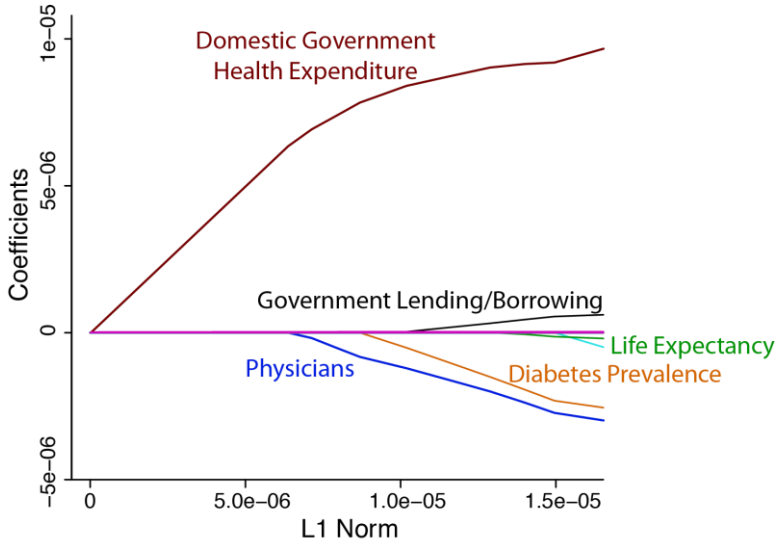

Fig. 1. For the outcome variable Y_2,, LASSO selected the following variables: Domestic general government health expenditure (per capita), the number of Physicians (per 1000 people), the prevalence of diabetes (%population), the government lending and borrowing (most recently available) and the life expectancy (at birth).

5. Limitations

At the current stage, our study suffers from some limitations. First, our study is static and uses a snapshot of the situation at 2 May 2020, while one of the most important factors of infectious diseases is their evolution. Second, it has been shown that underreporting can influence the severity of the pandemic [13,8,11,16], and therefore correct estimates of underreporting might

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

modify the results of our models. Finally, our models do not implement an age structure as [10], and spatio-temporal dependencies in the Y's as in [8].

6. Conclusions

Our study demonstrates the insights that can be provided by the inclusion of socio-economic factors into epidemiological models of the COVID-19 epidemic. Our results suggest the importance of focusing attention and efforts on developing countries where the healthcare system might not be equipped to deal with large numbers of COVID-19 cases before the availability of a vaccine. Further, environmental factors, such as pollution, need to be taken into account and may be a key focus of future epidemiological studies. Finally, our findings support social distancing measures and travel restrictions, but caution against long-term policy changes regarding immigration.

7. Acknowledgements

KLF and AMS would like to thank their families for their constant support.

References

[1] M. J. AZUR, E. A. STUART, C. FRANGAKIS, J. L LEAF: Multiple imputation by chained equations: what is it and how does it work? *Int. J. Methods Psychiatr. Res.*, **20**, 40–49 (2011).
[2] G. CHOWELL, C. VIBOUD, L. SIMONSEN, S. MOGHADAS: Characterizing the reproduction number of epidemics with early sub-exponential growth dynamics. *Journal of the Royal Society Interface*, **13**, 20160659 (2016).
[3] O. DIEKMANN, H. HEESTERBEEK, T. BRITTON: *Mathematical Tools for Understanding Infectious Disease Dynamics*. Princeton University Press, 1st edition, (2013).
[4] 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, S. FERRIER: Opinion: Sustainable development must account for pandemic risk. *Proceedings of the National Academy of Sciences*, **117**, 3888–3892 (2020).
[5] N. M. FERGUSON, D. LAYDON, G. NEDJATI-GILANI, N. IMAI, K. AINSLIE, M. BAGUELIN, S. BHATIA, A. BOONYASIRI, Z. CUCUNUBA, 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, 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*, 1–20 (2020).
[6] G. GIORDANO, R. BLANCHINI, R. BRUNO, P. COLANERI, A. D. FILIPPO, A. D. MATTEO, M. COLANERI: Modelling the covid-19 epidemic and implementation of population-wide interventions in Italy. *Nature Medicine*, 1–32 (2020).
[7] Global Footprint Network dataset retrieved from http://data.footprintnetwork.org/#/
[8] J. HILTON, M. J. KEELING: Estimation of country-level basic reproductive ratios for novel coronavirus (covid-19) using synthetic contact matrices. *medRxiv*, 1–7 (2020).
[9] International Monetary Fund dataset retrieved from https://www.imf.org/en/data
[10] J. P. IOANNIDIS, C. AXFORS, D. G. CONTOPOULOS-IOANNIDIS: Population-level covid-19 mortality risk for non-elderly individuals overall and for non-elderly individuals without underlying diseases in pandemic epicenters. *medRxiv*, 1–32 (2020).

Technium Social Sciences Journal
Vol. 10, 637-644, August 2020
ISSN: 2668-7798
www.techniumscience.com

[11] K. M. JAGODNIK, F. RAY, F. M. GIORGI, A. LACHMANN: Correcting under-reported covid-19 case numbers: estimating the true scale of the pandemic. *medRxiv*, 1–6 (2020).
[12] G. JAMES, D. WITTEN, T. HASTIE, R. TIBSHIRANI: *Introduction to Statistical Learning*. Springer, (2013).
[13] R. LI, X. PEI, B. CHEN, Y. SONG, T. ZHANG, W. YANG, J. SHAMAN: Substantial undocumented infection facilitates the rapid dissemination of novel coronavirus (sars-cov2). *Science*, **368**, 489–493 (2020).
[14] E. M. MIKHAEL, A. A. AL-JUMAILI: Can developing countries alone face coronavirus? The Iraqi situation. *Public Health in Practice*, 100004 (2020).
[15] Our World In Data, dataset retrieved from http://ourworldindata.org/
[16] D. P. SHUTTA, C. A. MANOREC, S. PANKAVICH, A. T. PORTER, S. Y. D. VALLE: Estimating the reproductive number, total outbreak size, and reporting rates for zika epidemics in South and Central America. *Epidemics*, **21**, 63–79 (2017).
[17] State of Global Air. dataset retrieved from https://www.stateofglobalair.org/engage
[18] V. STOJKOSKI, Z. UTKOVSKI., P. JOLAKOSKI, D. TEVDOSKI, L. KOCAREV: The socio-economic determinants of the coronavirus disease (covid-19) pandemic. *medRxiv*, 1–22 (2020).
[19] H. TIAN, Y. LIU, Y. LI, Y., C.-H. WU, ,B. Chen, M. U. G. KRAEMER., B. LI, J. CAI, B. XU, B., Q. YANG, B. WANG, P. YANG, Y. CUI, Y. SONG, P. ZHANG, Q. WANG, O. N. BJORNSTAD, R. YANG, B. T. GRENFELL, O. G. PYBUS, 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).
[20] R. TIBSHIRANI: Regression Shrinkage and Selection via the lasso. *Journal of the Royal Statistical Society. Series B (methodological)*. Wiley. **58**, 267–88 (1996).
[21] A. TUITE, D. N. FISMAN, A. L. GREER: Mathematical modeling of covid-19 transmission and mitigation strategies in the population of Ontario, Canada. *medRxiv*, 1–23 (2020).
[22] United Nations dataset retrieved from
https://population.un.org/Household/index.html#/countries/ 840
[23] S. VAN BUUREN, K. GROOTHUIS: mice: Multivariate imputation by chained equations in r. *Journal of Statistical Software*, **45**, 1–67 (2011).
[24] World Bank Open Data. dataset retrieved from https://data.worldbank.org/
[25] World Health Organization. Rolling updates on coronavirus disease (covid-19). https://www.who.int/ emergencies/diseases/novel-coronavirus-2019/events-as-they-happen