Inferring the main drivers of SARS-CoV-2 transmissibility

Marko Djordjevic[1,*], Igor Salom[2], Sofija Markovic[1], Andjela Rodic[1], Ognjen Milicevic[3], Magdalena Djordjevic[2]

[1]Quantitative Biology Group, Institute of Physiology and Biochemistry, Faculty of Biology, University of Belgrade, Serbia

[2]Institute of Physics Belgrade, National Institute of the Republic of Serbia, University of Belgrade, Serbia

[3]Department for Medical Statistics and Informatics, School of Medicine, University of Belgrade, Serbia

*** Correspondence:**

Marko Djordjevic, e-mail: dmarko@bio.bg.ac.rs

Keywords: COVID-19 environmental dependence, disease spread risk factors, basic reproduction number, principal component analysis, regression analysis, feature selection

Abstract

Identifying the main environmental drivers of SARS-CoV-2 transmissibility in the population is crucial for understanding current and potential future outbursts of COVID-19 and other infectious diseases. To address this problem, we concentrate on basic reproduction number R_0, which is not sensitive to testing coverage and represents transmissibility in an absence of social distancing and in a completely susceptible population. While many variables may potentially influence R_0, a high correlation between these variables may obscure the result interpretation. Consequently, we combine Principal Component Analysis with feature selection methods from several regression-based approaches to identify the main demographic and meteorological drivers behind R_0. We robustly obtain that country's wealth/development (GDP per capita or Human Development Index) is by far the most important R_0 predictor, probably being a good proxy for the overall contact frequency in a population. This main effect is modulated by built-up area per capita (crowdedness in indoor space), onset of infection (likely related to increased awareness of infection risks), net migration, unhealthy living lifestyle/conditions including pollution, seasonality, and possibly BCG vaccination prevalence. Also, we show that several variables that significantly correlate with transmissibility do not directly influence R_0 or affect it differently than suggested by naïve analysis.

1 Introduction

Despite the unprecedented worldwide campaign of mass immunization, due to the relatively slow vaccine rollout and to the appearance of new, more contagious (Tegally et al., 2020), and maybe even more deadly SARS-CoV-2 strains (Mallapaty, 2021), COVID-19 still takes its toll on human lives, stifles the world economy, and forces the majority of countries to keep unpopular lockdowns. In the absence of a prompt solution to the first pandemic of the century, the goal to identify the main environmental and demographic parameters that influence the dynamics of infection transmission remains as important as ever.

We recently published a comprehensive study of the correlation of 42 different demographic and weather parameters with COVID-19 basic reproduction number R_0 across 118 world countries (Salom et al., 2021). R_0 is a well-established epidemiological measure of virus transmissibility, which has a major advantage of being independent on the testing policy/capacity, and on the intervention measures that can be highly variable (and almost impossible to consistently control) between different countries (Salom et al., 2021). In (Salom et al., 2021), we selected all the countries that exhibited regular exponential growth in the case numbers before the introduction of intervention measures (Djordjevic et al., 2021), from which their R_0 values can be reliably extracted. Tracking a wide range of countries allows achieving a maximal variability in the dataset, i.e., a maximal possible range in the values of analyzed variables, as another advantage of this study. This generated dataset will be used as a starting point in this work.

While (Salom et al., 2021) covered a broad scope of variables and countries, it focused on establishing pairwise correlations between R_0 and each of the studied factors, ignoring the fact that many of these variables are highly mutually correlated. This is most obvious in the case of the weather parameters such as e.g. temperature and UV radiation (which both reflect the local climate in a similar way and follow comparable seasonal trends), but also in the case of many demographic parameters, e.g. the strong positive correlation between the Human Development Index (HDI) and cholesterol levels. Based on pairwise correlations alone, it is thus hard to estimate which of these variables might be truly influencing the spread of the disease, to what extent, and in which direction. To achieve this, the number of variables necessary to explain the virus transmissibility needs to be reduced to only a few without losing predictiveness. However, this is not the only challenge, because of variable redundancy. In particular, one may select different combinations of variables accounting together for a similar proportion of variance in the virus transmissibility, which seems to be a dead-end (Notari and Torrieri, 2020). There is consequently a challenge to narrow down the possibilities and illuminate important contributions of the seemingly small differences between highly correlated variables. Noticeably, while numerous studies examined the correlations of several selected (Lin et al., 2020; Ran et al., 2020; Xie et al., 2020) or many different (Li et al., 2020; Hassan et al., 2021; Salom et al., 2021) sociodemographic and meteorological factors with the magnitude of the COVID-19 epidemic, only few studies tried to select a handful of key factors whose combination can explain a large portion of the variance between regions (Allel et al., 2020; Coccia, 2020; Gupta and Gharehgozli, 2020; Notari and Torrieri, 2020). Even a smaller number of studies included data from multiple countries (Allel et al., 2020; Notari and Torrieri, 2020).

The main idea of this study is to develop a *novel* approach to robustly identify the most important predictors of R_0. The development of such an approach will *i*) provide a straightforward solution to the known problem of selecting important among the highly correlated variables, *ii*) enable a better understanding of which environmental and demographic variables may dominantly and/or

independently influence the progression of the COVID-19 epidemics, and what is the direction of this influence. To achieve these goals, the study is organized as follows:

1. The variables are first naturally split into two groups. The first group comprises 6 meteorological parameters, sampled and averaged (for each country) during the initial stage of the local epidemic outbreak: air temperature (T), precipitation (PC), specific humidity (H), ultra-violet radiation index (UV), air pressure (P), and wind speed (WS). Eighteen (broadly-speaking) demographic parameters form the second group: human development index (HDI), percentage of the urban population (UP), gross domestic product per capita (GDP), amount of the built-up area per person (BUAPC), percentage of refugees (RE), net migration (i.e., the number of immigrants minus emigrants, I-E), infant mortality (IM), median age (MA), long-term average of PM2.5 pollution (PM), prevalence and severity of COVID-19 relevant chronic diseases in the population (CD), average blood cholesterol level (CH), the prevalence of raised blood pressure (RBP), the prevalence of obesity (OB), the prevalence of insufficient physical activity among adults (IN), BCG immunization coverage (BCG), alcohol consumption per capita (ALC), smoking prevalence (SM), and the delay of the epidemic onset (ON).
2. Due to strong mutual correlations between parameters within each group (as well as across the groups, but at a lower extent), the principal component analysis (PCA) will be performed on each of the groups (Jolliffe, 2002). This step will allow us to notably reduce the dimensionality of the problem, i.e., proceed to work with a smaller number of (mostly) uncorrelated variables. Such dimensionality reduction will significantly simplify the further analysis and improve the reliability of the results.
3. The linear regression analysis will next be performed in four independent ways, ranging from our custom-developed to more formal regression-based approaches, to select important variables. In our custom-developed approach, multiple linear regressions are applied, first separately to demographic and meteorological principal components (PCs), to narrow down the number of relevant PCs within each of the two groups, before doing overall linear regression with the remaining PCs to assess their importance in explaining R_0. A major advantage of such analysis is in an intuitive understanding of the data structure and its relation to R_0. This analysis is next independently redone by more formal feature selection methods, commonly employed in bioinformatics and systems biology: Stepwise regression and regressions utilizing both regularization and variable selection - LASSO (Least Absolute Selection and Shrinkage Operator) and Elastic net (Tibshirani, 1996; Zou and Hastie, 2005; Hastie et al., 2009). Such comprehensive analysis will ensure the consistency and robustness of the reported results.
4. Finally, an intuitive interpretation of the obtained results will be presented. This will permit a much more specific understanding of COVID-19 transmissibility, by focusing on the main driving factors behind the disease spread in the population.

2 Methods

2.1 Data collection

Data for demographic and meteorological parameters were assembled as described in (Salom et al., 2021). Briefly, the data correspond to six meteorological and eighteen demographic variables outlined above. The differences between this dataset and the one used in (Salom et al., 2021) is the following: IMS (Social security and health insurance coverage), Prevalence of ABO and Rhesus blood groups, and Ambient levels of different pollutants (NO_2, SO_2, CO, PM2.5, PM10) are not used in this analysis, as they contain too many missing values. Instead of the pollutant levels measured from air pollution monitoring stations during the epidemic's exponential growth (available for only ~40 countries) we

use the yearly average PM2.5 pollutant levels in 2017 (World Bank, 2020b). Also, we consider GDP per capita (GDPpc), taken from (World Bank, 2020a) as a more direct (average) indicator of a country's economic wealth/productivity.

Basic reproduction number (R_0), i.e., a measure of SARS-CoV-2 transmissibility in a fully susceptible population and in the absence of intervention measures (social distancing, quarantine), was also taken from (Salom et al., 2021), where it was inferred from non-linear dynamics modeling. Overall, demographic data, meteorological data, and basic reproductive numbers were assembled for 118 different countries from which we could reliably infer R_0. Missing values in the demographic data (which were sparse for the used variables) were substituted by median values of the respective variables; there are no missing values in the meteorological data.

2.2 Data preparation

Several variables, particularly among demographic data, show a significant deviation from normality when visually inspected. Such deviations generate large outliers and would significantly impact the necessary normality of the model error residuals. We consequently transform the data where necessary, to make the resulting distributions closer to normal, by using standard transformations that reduce the right and left skewness. The strength of the applied transformations (e.g., square root, cubic root, or log) is chosen so that skewness of the transformed distribution is as close to zero as possible. The table with all applied transformations is provided below:

Variable	Transformation
BUAPC	$(x - \min(x))^{1/3}$
UP	x^2
IM	$\log(x)$
GDPpc	$\log(x)$
HDI	$(\max(x) - x)^{1/2}$
I-E	$(\max(x) - x)^{1/2}$
RE	$\log(x)$
CH	$(\max(x) - x)^{1/2}$
OB	$(\max(x) - x)^{1/2}$
CD	$x^{1/3}$
IN	$\log(\max(x) - x)$
BCG	$(\max(x) - x)^{1/2}$
ON	$\log(x)$
PL	$\log(x)$
WS	$\log(x)$
P	$x^{1/3}$
R_0	$\log(x)$

After transformations, the remaining (now sparse) outliers were removed by substituting them with the median of each variable; the outliers were identified as having more than three scaled median absolute deviation (MAD) from the (transformed) variable median. Each transformed variable whose direction was changed by the transformation was taken with a minus sign, so that the original and the transformed variable are oriented in the same direction, allowing for easier result interpretation.

2.3 Principal components analysis

The dimensionality of the transformed data was reduced and the data decorrelated through PCA (Jolliffe, 2002). PCA was done separately for demographic and meteorological variables to allow for a more straightforward interpretation of the obtained PCs. Since different variables are expressed in different units and correspond to diverse scales, each variable in the dataset was standardized (the mean subtracted and divided by the standard deviation) before PCA. For both datasets, we retained as many PCs (starting from the most dominant one) as needed to (cumulatively) explain >85% of the data variance. It was inspected that PCs reasonably follow a normal distribution (as expected, based on the transformation of the original variables). Few remaining outliers for PCs were then substituted by medians. For easier interpretation of PCs and their contribution to R_0, each PC was oriented in the same direction as the variable with which it has a maximal magnitude of Pearson correlation (i.e., the sign of the PC was flipped when needed, to render the positive sign of this correlation).

2.4 Custom regression analysis

Multiple linear regression (PC regression) was done first with only demographic PCs (Hastie et al., 2009). Only linear terms were included in the regression to allow straightforward interpretation, i.e., selection of PCs that significantly affect R_0. Significant PCs were selected as those appearing in the regression with $P<0.05$, where the significance in the regression was estimated in the standard way (through F-statistics) (Alexopoulos, 2010). The same, regression was then repeated with only meteorological PCs, and those significant in explaining R_0 were retained. Finally, multiple linear regression was performed with all retained demographic and meteorological PCs. The significant PCs from this last step were recognized as PCs relevant for R_0 explanation. Before regression, each PC was standardized so that coefficients obtained in the regression provided a measure of the variable importance in explaining R_0. For both the custom analysis and stepwise regression, OLS (Ordinary Least Squares) were used as the regression metrics.

2.5 Stepwise regression

Stepwise regression was used to select PCs that significantly affect R_0. In Stepwise regression, as well as in LASSO and Elastic net described below, all PCs (demographic and meteorological) were included in the regression. Briefly, starting from a constant model, at each step a term is added to the model if its significance (calculated with F-statistics) meets the condition $P<0.05$ (Pope and Webster, 1972). Only linear terms are added to the model (i.e., interaction and quadratic terms are not considered) to allow for straightforward interpretation which PCs significantly affect R_0. All PCs are standardized before regression so that contributions of the terms (PCs) in the model can be assessed by the magnitude of the regression coefficient.

2.6 LASSO regression

L1 regularization was implemented through LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani, 1996; Hastie et al., 2009). As needed with the LASSO regularization, all PCs were standardized before regression, which also allowed direct comparison of the coefficients obtained by the regression. The value λ in LASSO was treated as the hyperparameter, i.e., $\lambda_{\min}$ value was determined through cross-validation, so that MSE (Mean Squared Error) on the testing set was minimal. A total of 100 λ values were put on the grid, corresponding to the geometric sequence, where the largest value produces all zero terms. Note that larger λ corresponds to sparser model, i.e., a smaller number of non-zero components in the regression, while the small λ limit corresponds to OLS regression. To obtain the maximally sparse model, $\lambda_{1\mathrm{SE}} = \lambda_{\min} + 1\mathrm{SE}$, where 1SE corresponds to the

standard error of MSE obtained by cross-validation, was used. 1000 cross-validations were performed, where in each repetition 20% of the data were randomly selected for the testing set, with the remainder used for training. All non-zero terms and the corresponding coefficients obtained through LASSO were reported.

2.7 Elastic net regression

A combination of L1 and L2 regularization was implemented through Elastic net regression (Zou and Hastie, 2005). Analogously to our LASSO analysis, i.e., as needed due to regularization, all PCs were standardized. In the regression, both α and λ were treated as hyperparameters, i.e., their optimal values were found by cross-validation. Cross-validation was repeated 1000 times, wherein each repetition testing and training sets were formed in the same way as for LASSO. α and λ values were put on a grid consisting of 100 α and 100 λ values. α values on the grid were chosen uniformly in the range [0,1] - α approaching zero corresponds to Ridge (L2) regression, and 1 corresponds to LASSO regression. For each α value, λ values were chosen as described for the LASSO regression. For each repetition of cross-validation, α and λ combination which leads to the minimal MSE was chosen. α and λ values in (α, λ) pairs from each cross-validation run were then standardized so that α and λ values are on the same scale and centered to the origin of the $\alpha - \lambda$ plane. $(\alpha_{\min}, \lambda_{\min})$ was then chosen as the (α, λ) point closest to the origin. With this $(\alpha_{\min}, \lambda_{\min})$ value the model was then retrained on the entire dataset. Similarly to LASSO, all non-zero terms and the corresponding regression coefficients were reported.

3 Results

3.1 Dimensionality reduction of the demographic dataset

PCA was first applied to the dataset consisting of 18 demographic and health factors for 118 countries. Cumulative data variance that is explained jointly by the first *n* PCs is shown in Figure 1A (with *n* represented on the x-axis). In particular, Figure 1A shows the first PC alone already accounts for 45% of the variance, while the first 9 PCs (PC1 – PC9), which we retain in further analysis, explain more than 85% (precisely, 89%).

To obtain a basic interpretation of these nine PCs, we related each PC with the original (transformed) variable it is most correlated with. The corresponding associations – with the values of correlations coefficients presented on the y-axis – are shown in Figure 1B (however, one should have in mind that some PCs are highly correlated with more than one original variable, as we discuss in more detail below). Among all principal components, the PC1 and the PC5 have the highest correlation coefficients (close to 1) with individual demographic factors – the HDI and the BCG immunization coverage, respectively. Moderately high correlation coefficients (~0.75) characterize the relations between the PC2 and the prevalence of smokers, and the PC3 and the percentage of refugees, while the coefficient values of ~0.5 were obtained for the correlations of the PC4, the PC6, the PC7, the PC8 and the PC9 with, respectively, the prevalence of obesity, the prevalence of insufficient physical activity, the amount of the built-up area per person, the percentage of refugees, and the epidemic onset.

In particular, the first PC, accounting alone for the largest portion of the variance in the demographic data, is almost perfectly correlated with the Human Development Index (Fig. 1C). On the other hand, the HDI variable itself strongly correlates with several other demographic variables (Fig. 1D), most prominently with per capita GDP, infant mortality, and cholesterol levels. As elaborated in the Discussion section, such extremely high correlations will eventually preclude us from differentiating between the separate effects of each of these variables on R_0. On the other hand, the prevalence of

obesity, the built-up area per person, and the epidemic onset are significantly correlated with the HDI (Fig. 1D), and thereby the PC1 (Fig. 1C), but they are markedly featured also in separate principal components (Fig. 1B), namely – the PCs 4, 7 and 9. This will help us to infer whether their specific, additional contributions to the variance in the data (apart from that along the PC1) impact the virus transmissibility.

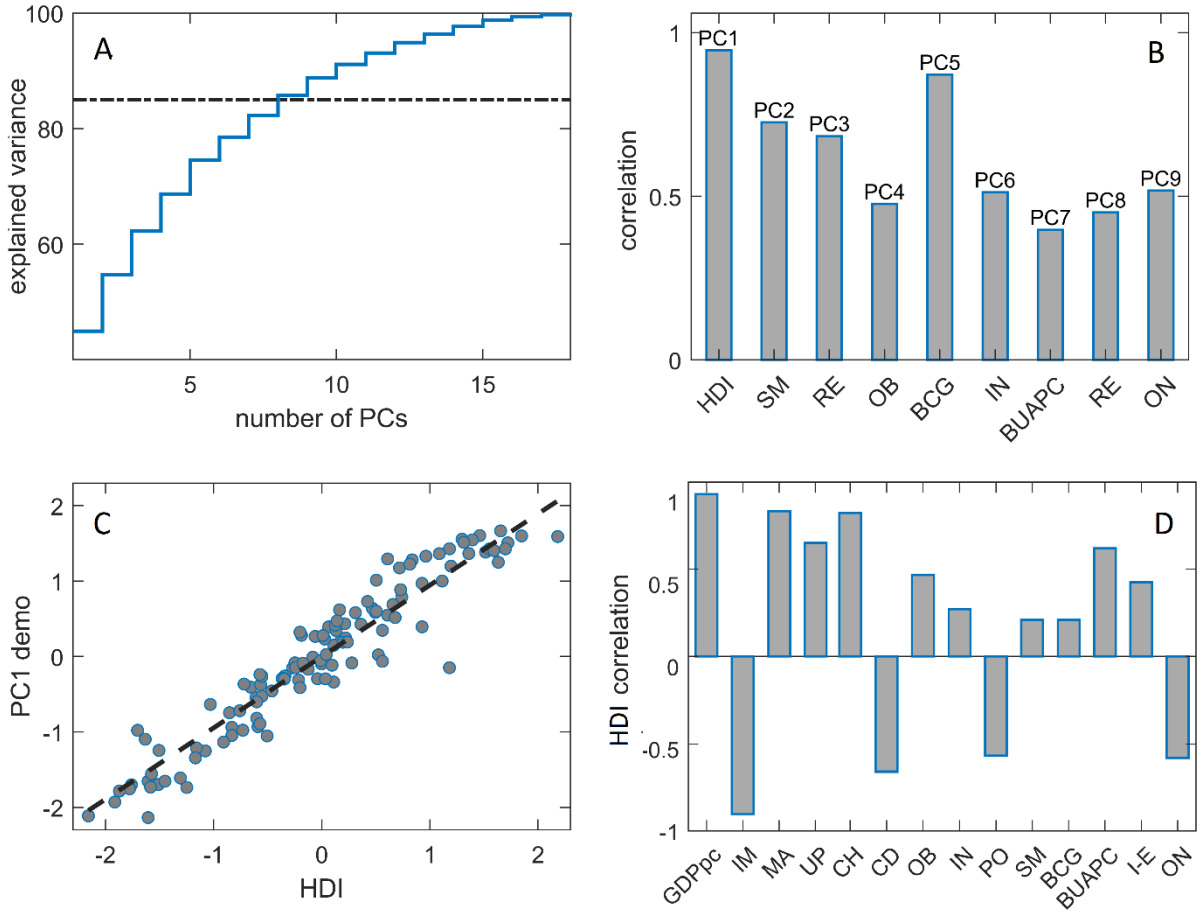

Figure 1. PCA for demographic data. **A)** Cumulative explained variance. **B)** Variables best correlated with demographic PCs. The label above and below each bar present, respectively, the demographic PC and the variable with which this PC has the highest correlation. **C)** Scatter plot PC1 vs HDI. **D)** Correlations of selected demographic variables with HDI.

3.2 Dimensionality reduction of the meteorological dataset

The dimensionality of the dataset consisting of 6 meteorological factors for 118 countries was reduced similarly as for the demographic dataset. PCA generated 6 uncorrelated, orthogonal principal components. Thereby, the first PC alone explains 62% of the variance, while the first three PCs (PC1-PC3) capture 95%, which is significantly above the targeted 85% of the total variance (Fig. 2A). Pairwise correlations showed that the retained three PCs have the highest correlations with the temperature, the wind speed, and the air pressure, respectively (Fig. 2B), where the correlation of PC1 with the temperature is close to 1 (Figs. 2B and 2C). There are also notable correlations of the temperature with humidity, the levels of UV radiation, and precipitation (Fig. 2D). PC1, therefore, presents seasonality, i.e. a set of mutually correlated meteorological variables which can be related to yearly weather changes. Consequently, PCA effectively separated the impacts of seasonality (PC1), the wind speed (through the PC2), and the air pressure (through the PC3). The variables determining the PC1 are also correlated with the HDI. These inter-dataset correlations are not resolved at this level

by our PCA and represent the trade-off that allows interpreting the PCs more easily within each of the two smaller, thematic groups of factors.

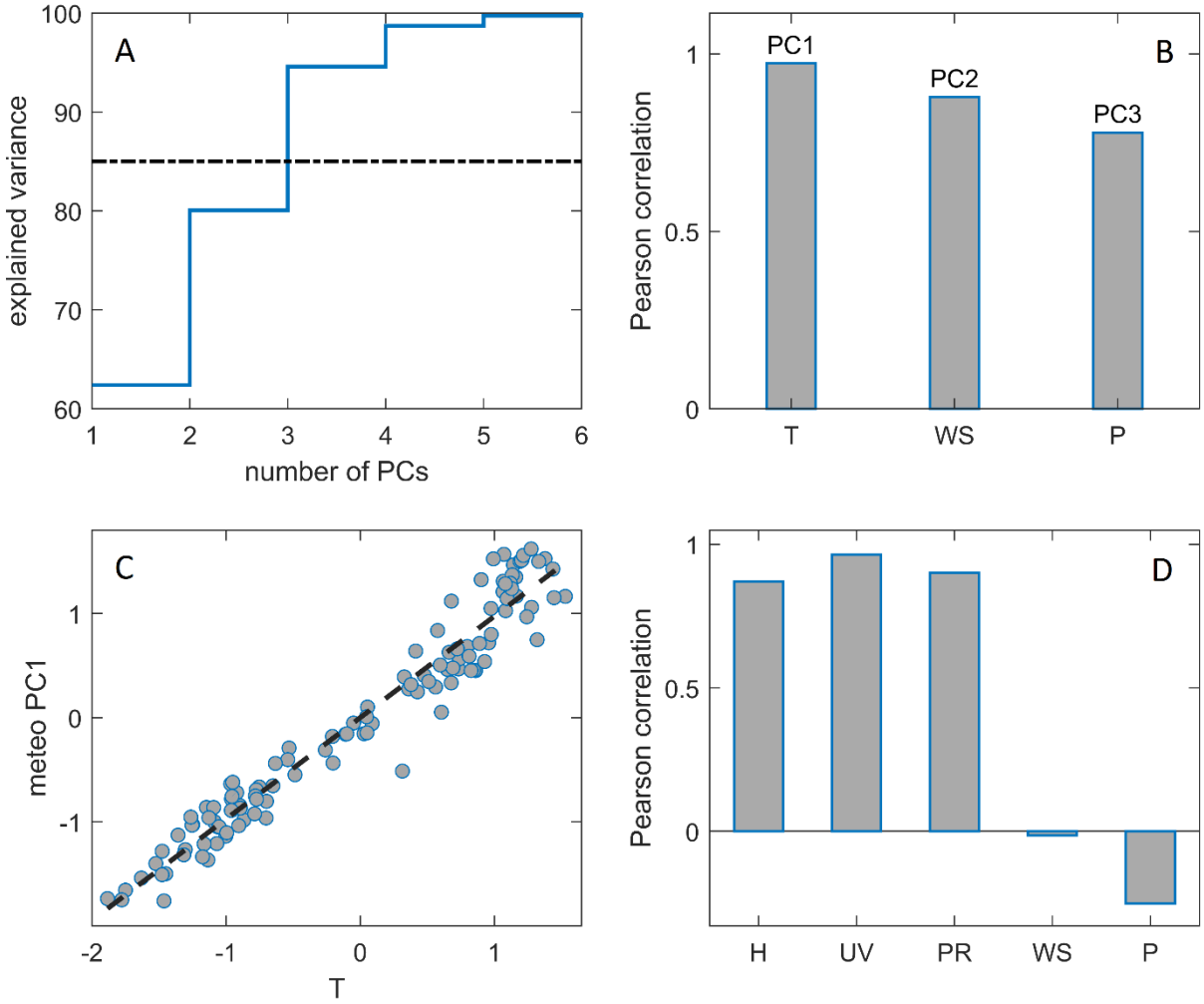

Figure 2. PCA for meteorological data. **A)** Cumulative explained variance. **B)** Variables best correlated with meteorological PCs. **C)** Scatter plot meteo PC1 vs temperature. **D)** Correlation of meteorological variables with temperature.

3.3 Linear regressions

After PCA, we applied the linear regression analysis using four different methods, as explained in Methods. The first, "custom" method included the additional step of "preselecting", i.e. further narrowing down the number of PCs that will enter the final regression analysis. The multiple linear regression, applied on the group of 9 demographic principal components, selected 1st, 4th, 7th and 9th component as the most relevant predictors of R_0 (the remaining 5 components appeared in the linear regression with p values above 0.05 threshold, and were consequently excluded from the further analysis). Analogously, the "preselection" of meteorological principal components singled out the 1st component as the only statistically relevant predictor of R_0 from this group. The multiple linear regression was then applied on these 5 selected PCs (4 demographical and 1 meteorological) and yielded a regression model with the corresponding linear coefficients represented in Figure 3A. Meteo PC1 component does not appear in the results of the custom method, due to the lack of statistical significance ($p>0.305$) in the final regression, so that according to our custom regression methodology, weather parameters do not significantly influence R_0. R_0 in this model is therefore determined by a

combination of demographic PC1, PC4, PC7, and PC9, where coefficients multiplying PC1 and PC4 are positive, while for PC7 and PC9 are negative. As can be inferred from the values represented in Fig 3A, the demographic PC1 has the most dominant influence on R_0 – a robustly obtained result throughout all 4 methods (see below).

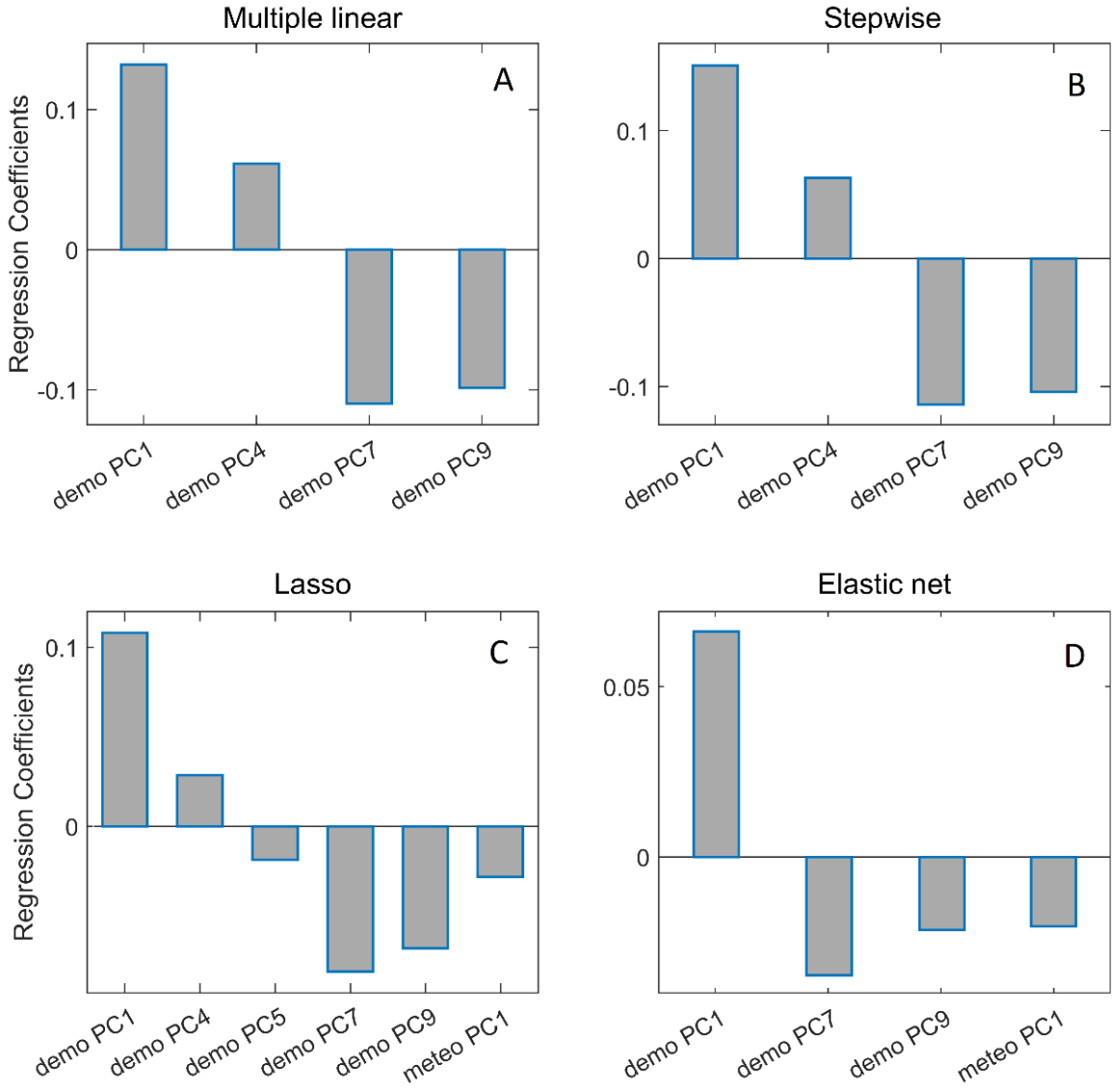

Figure 3. Results of: **A)** multiple linear regression ("custom") method, **B)** Stepwise regression, **C)** LASSO regression and **D)** Elastic net regression. Bar charts represent the values of regression coefficients for each of the PCs selected by the method.

We have already related each of these four PCs with the dominantly correlated variable (Figure 1B), but a more detailed interpretation of the results is obtained if all significant correlations (not just the dominant one) are taken into account. In addition to the very high correlation with HDI, demographic PC1 is also highly positively correlated with GDP, cholesterol levels, median age, and percentage of the urban population, while it is highly negatively correlated with infant mortality and the prevalence of chronic diseases (Figure 4A). Such strong correlations with HDI, GDP, IM, MA, and UP show that this component indeed expresses an overall, both social and financial, prosperity of the country (which seemingly also goes hand in hand with high average cholesterol levels and low prevalence of COVID-19 relevant chronic diseases). Similarly, by considering the correlations of demo PC4 with all

demographic variables, we see that this component is significantly positively correlated not only with obesity but also with smoking, physical inactivity, and air pollution (Figure 4B) – in other words, with major indicators of an unhealthy lifestyle and living conditions. Apart from its correlation with the BUAPC parameter, the component demo PC7 is also significantly positively correlated with net migration (Figure 4C). In the case of the demo PC9 component, its only significant correlation is with the onset variable. Results of the custom method can therefore be summarized as follows: the country's prosperity, as well as unhealthy living conditions and lifestyle, tend to increase the value of R_0, while the larger built-up area per person and the later epidemic outbreak tend to slow the spread of the disease. Also, the results seem to indicate – via demo PC7 component – a surprising diminishing effect of the net migration on the rate of epidemic progress (though the sign of this variable may not be easy to interpret, as the net migration is a difference of two quantities).

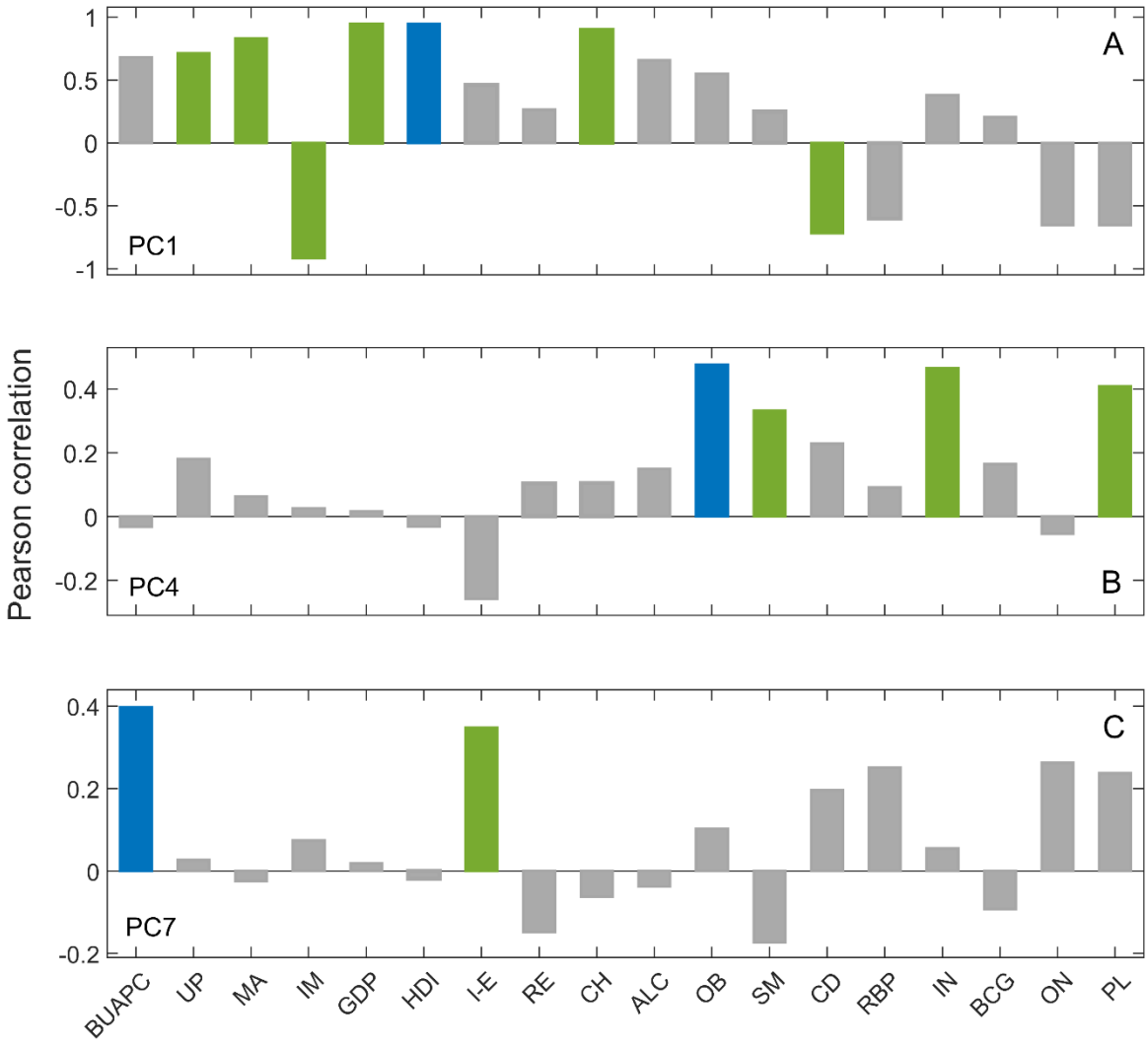

Figure 4: Pearson correlation coefficients between principal components and demographic variables for **A)** demo PC1, **B)** demo PC4, and **C)** demo PC7.

Equivalently to Figure 3A, Figures 3B, 3C, and 3D represent the results of, respectively, Stepwise, LASSO, and Elastic net regression. Results (and the corresponding graph) of the Stepwise method almost coincide with the results of our custom method – in spite that in the Stepwise regression (as well as in LASSO and Elastic net methods) there is no intermediate "preselection" step.

LASSO results, shown in Figure 3C, find two additional PCs as relevant: demo PC5 and meteo PC1 (in addition to demo PC1, demo PC4, demo PC7, and demo PC9). The component demo PC5, appearing in LASSO results with a small negative coefficient, is significantly correlated only with the BCG variable, hinting at possible beneficial effects of BCG vaccination. Meteorological principal component meteo PC1 reflects seasonality (see above). Thus, overall, in addition to supporting the conclusions of the custom and stepwise methods, the LASSO method also implicates a significance of seasonality changes, and to some extent BCG vaccination, in reducing the rate of SARS-CoV2 spread.

The results of the Elastic net method, shown in Figure 3D, are again a bit more restrictive. While further bolstering our confidence in the importance of demo PC1, demo PC7, and demo PC9, these results also reinforce that the seasonal weather variables influence the COVID-19 epidemic (in agreement with the LASSO method) but, for the first time, we do not find an indication of the relevance of the unhealthy lifestyle and living conditions – as revealed by the absence of demo PC4 component in Figure 3D.

Finally, as much as the PCs appearing in Figure 3 are important, the absence of the remaining PCs in the results can be of comparative significance for some of our conclusions. For example, we note that PCs highly correlated with the urban population, alcohol consumption, and chronic diseases do not show up as relevant in any of the methods used. While it is true that these variables are moderately correlated with demo PC1, absence in the results of additional PCs tied with these variables supports the view that these variables are not directly influencing R_0 value, but only via indirect relation to the country's prosperity.

4 Discussion

Our goal was to identify the most predictive factors influencing the risk of the SARS-CoV-2 virus spreading in a population in the absence of any epidemic mitigation measures. Since many potentially relevant factors strongly correlate with each other, we divided them into two groups –meteorological and sociodemographic – and applied the Principal Component Analysis to the variables in each group. In this way, we were able to decorrelate variables within each group, while still retaining intuitive interpretation for the new variables (demographic and meteorological PCs) used in further analysis. Dimensionality reduction and predictor decorrelation through PCA was then combined with different variable selection and regularization techniques, to select PCs that are most predictive of R_0 for COVID-19 epidemics. Examining correlations of these PCs with the original variables allowed pinpointing the main drivers of COVID-19 transmissibility. This approach is to our knowledge unique in the COVID-19 research literature, and reminiscent of the analysis of complex data in systems biology and bioinformatics.

Three principal components are robustly selected as the most important predictors by all the methods. Of these, the prosperity of the country has the most significant influence on R_0: the spread of the epidemic is faster in economically more developed countries. Specifically, this is the most dominant PC from the demographic group of variables, which is by far most important in explaining R_0, and very strongly correlated with HDI (Pearson's correlation coefficient r=0.95) and GDP (r= 0.94) – therefore effectively reflecting prosperity and wealth. The second PC is dominantly related to the built-up area per person (BUAPC), and the third with the epidemic onset, where the increase of these reduces the infection spread. We also robustly obtained (by three out of four methods) that unhealthy living conditions and lifestyle – i.e., the PC dominantly (and consistently positively) correlated with obesity, physical inactivity, smoking, and air pollution – is another important factor that exacerbates the epidemic. Seasonality, represented by the group of four weather conditions all significantly correlated with temperature, was selected by two independent methods including, importantly the Elastic net, which is well adapted to selecting among correlated variables (Zou and Hastie, 2005; Hastie et al.,

2009) - note that correlations between meteorological and demographic PCs were not abolished by our approach. The PC dominantly correlated with BCG immunization appears only in LASSO regression.

4.1 *High economic development as the main predictor of COVID-19 transmissibility*

As noted above, we consistently obtained that the first demographic PC is the most important predictor of R_0. HDI (alternatively, GDPpc) shows the highest correlation with this PC, which singles out this variable as the main index quantifying the virus transmissibility risk. Higher HDI leads to a higher rate of social contacts and more intense population mixing, as high HDI is strongly associated with high GDPpc implying intensive economic activity, trade, and transportation, including large-distance flights (Allel et al., 2020; Gangemi et al., 2020). Thus, much higher contact frequency in societies with higher HDI is likely the main cause behind the dominant role of the first demographic PC in explaining R_0.

An important advantage of our approach is that it is based on the analysis of R_0, rather than other measures used as transmissibility proxies. The most commonly used measure, confirmed case counts, strongly depends on the number of performed tests, which is generally much higher in high-GDPpc countries, so the analysis would become strongly influenced by testing policies. For example, in (Allel et al., 2020) the importance of HDI for predicting cumulative case counts was noted. However, this perceived effect may be due to the lack of testing in lower-income countries (Notari, 2021), rather than genuine HDI influence. Our results are, on the other hand, insensitive to the testing capacity differences, since our R_0 estimation procedure relies on the slope of the case growth curve in the distinct early exponential phase (Djordjevic et al., 2021), which requires only that the testing is performed consistently during the relatively short examined period (Salom et al., 2021). Therefore, our analysis indeed strongly suggests that HDI/GDPpc are the main/genuine predictors of COVID-19 spread in the population.

4.2 *Demographic factors significantly correlated with HDI*

Many correlations previously reported between SARS-CoV-2 transmissibility and various weather, sociodemographic, and health factors [see e.g. (Li et al., 2020; Salom et al., 2021)] may be captured by HDI. From our results, one can note that several demographic factors significantly correlate with both HDI/GDPpc and the first demographic PC, but are not noticeably related with other demographic PCs (4,5,7,9) that significantly contribute to R_0. These demographic factors can be further divided into two groups using the correlation of BUAPC with HDI as the reference. The percentage of the urban population, the prevalence of alcohol consumption, and chronic diseases, which have similar (just somewhat higher) correlations with HDI compared to BUAPC, comprise the first group. Their absence from the independent PCs significantly related with R_0, in contrast to BUAPC which prominently appears in the demographic PC7, indicates that they do not have independent effects on R_0. Consequently, their significant correlation with R_0 (Salom et al., 2021) is very likely due to their generic correlation with HDI, rather than a consequence of the independent effect that they exhibit on R_0. This result is especially interesting for the percentage of the urban population, whose relation with R_0 is sometimes taken for granted (Carozzi, 2020). It also explains the previously obtained negative correlation of the prevalence of chronic diseases with R_0, where one might expect the opposite, as it is generally known that people with chronic diseases are seriously affected by COVID-19 (Zheng et al., 2020). We can now claim that this result is due to a generically lower incidence of chronic diseases in more developed countries (i.e., due to their significant negative correlation with HDI), rather than a direct effect on R_0.

The net economic immigration (the difference between immigrants and emigrants), population median age, infant mortality, and the average blood cholesterol level, comprising the second group, also have a significant positive correlation with the first demographic PC. However, in distinction to the aforementioned three factors, their correlation with HDI is very high, i.e., visibly higher compared to the correlation of BUAPC with HDI. So, even though they do not appear in demographic PCs that significantly contribute to R_0 other than PC1, we cannot make any reliable conclusion about their direct effect on R_0 based on our analysis. It is therefore relevant to discuss evidence from other sources, i.e., possible mechanisms that can distinguish their direct influence on R_0. Regarding infant mortality, a mechanism of its direct contribution to R_0 is hard to imagine, so its involvement in PC1, and high negative correlation with R_0, is almost certainly an indirect consequence of this variable being a proxy of HDI (Ruiz et al., 2015). On the other hand, the median age and the blood cholesterol level are real contenders for direct R_0 modifiers, as mechanisms for their contribution to COVID-19 transmissibility have been proposed. Aging is generally associated with the weakening of the immune response to infectious diseases making the elderly more susceptible to the viruses like the SARS-CoV-2 (Pawelec and Larbi, 2008). Additionally, many of them due to some chronic diseases take ACE inhibitors and angiotensin-receptor blockers which cause an increased expression of ACE2 serving as a receptor for the SARS-CoV-2 virus entry (Shahid et al., 2020). Their residing in care-homes, which is particularly common in high-income countries, also well suits the spreading of the infection (Kapitsinis, 2020). Similarly, high cholesterol levels can increase susceptibility to the infection by SARS-CoV-2 through systemic adverse effects on the immune and inflammatory responses, but also through direct implication in the virus life cycle, especially at the level of its endocytosis. To that end, statins, blocking cholesterol synthesis, were proposed for usage in COVID-19 treatment, which is supported by studies showing that previous statin usage is associated with a milder pneumonia outcome in the case of several other viral infections (Frost et al., 2007; Schmidt et al., 2020).

4.3 *Independent COVID-19 transmissibility predictors*

All the demographic variables discussed in the previous subsection show a rather strong correlation with the first demographic PC but are not involved with other significant demographic PCs (4,5,7,9). These PCs are by construction independent (decorrelated) from PC1. Variables associated with these PCs can be interpreted as effects on R_0 independent from those related to PC1. These variables then importantly identify corrections to the main effect of HDI/GDPpc. Specifically, these are indoor area available to an individual and the net immigration (demographic PC7), the delay in the epidemic onset with respect to February 15th associated with more awareness of the virus threat (demographic PC9), the prevalence of unhealthy lifestyle and environment (demographic PC4), and the weather seasonality (meteorological PC1).

The slower spread of the virus with a larger built-up area per capita, as an independent and significant R_0 predictor, is an interesting and new result, though intuitively plausible. It can be understood as having a less crowded indoor space (where the virus transmission dominantly happens) so that people are less exposed to each other and the virus. For example, both the population density and R_0 on the Diamond Princess cruise ship were estimated as four times greater than those in Wuhan (Rocklöv and Sjödin, 2020). On the other hand, a correlation of the virus transmissibility with the large territory population density is weakly established in the literature, whereby it seems that one should rather seek a correlation with a local population density, directly determining the number of contacts that an individual can make (Garland et al., 2020).

A positive contribution to the transmissibility is also made by the principal component strongly correlated with the onset variable, representing the number of days from February 15th to the

epidemic's start in a particular country. The importance of the delay in the epidemic onset may be due to the psychological effect of hearing the news about the spread of COVID-19 in other countries (Khajanchi et al., 2020). Namely, the longer the epidemic was growing outside of a particular country, the larger impact this had on its people to change their usual behavior to prevent the infection, which could slow down the virus transmission even before the introduction of the official intervention measures (Salom et al., 2021).

Another distinguished principal component appears to encompass multiple indicators of an unhealthy lifestyle and environment – specifically, the prevalence of obesity, physical inactivity, and smoking, together with the level of air pollution. We obtained that all these factors promote virus transmission. It is well established that they can impair immune function and adversely affect different organ systems. Furthermore, their association with mechanisms specifically facilitating the infection by the SARS-CoV-2 virus has been proposed (Domingo and Rovira, 2020; Heidari-Beni and Kelishadi, 2020; Haddad et al., 2021).

Two more PCs are strongly determined by temperature (and/or three other highly related weather factors) and the prevalence of BCG vaccinated children, respectively. Although not selected by all the methods, the weather component seems important as it was chosen by the Elastic net algorithm (in addition to LASSO), which is specifically designed to deal with (highly) correlated variables, and yet it did not exclude this PC despite its correlation with the first demographic PC. Moreover, a decrease of the transmissibility with the temperature increase appears as a robust result in COVID-19 literature, although conflicting conclusions are also present (Srivastava, 2021). Higher temperatures may shorten the period of virus viability in aerosols, enhance the immune system functioning, and/or impact the time that people spend together in poorly ventilated indoor spaces (Notari, 2021). Since temperature is highly positively correlated with the intensity of UV radiation, humidity, and the level of precipitation, we cannot exclude the possibility that some of these other factors are in a significant causal relationship with virus transmissibility. Importantly, some experimental findings support the inactivating effects of high temperature, humidity, and UV radiation on SARS-CoV-2 and related viruses (Casanova et al., 2010; Chan et al., 2011; Heilingloh et al., 2020; Sagripanti and Lytle, 2020; van Doremalen et al., 2020). Anyhow, our results suggest the dependence of virus transmissibility on seasonal weather variations.

Regarding the last demographic principal component, it occurred as important only in LASSO regression, but it closely follows the extent of BCG vaccination, which is known to provide some protection against various respiratory tract infections through the induction of the trained immunity (O'Neill and Netea, 2020), so BCG immunization may significantly influence the SARS-CoV-2 spread, although, according to our results, to a lesser extent than the other discussed factors.

4.4 Differences to pairwise correlation analysis

Our study is also an example of how assessing the effect of one factor while controlling for the presence of other relevant variables can change the obtained conclusions. We will illustrate this with four examples, where we obtained qualitatively different conclusions, compared to single-variable correlation analysis (Salom et al., 2021): built-up area per capita (BUAPC), net migration, air pollution, and raised blood pressure.

BUAPC showed an absence of a significant correlation with R_0 (Salom et al., 2021), which is due to the canceling of two effects. The first is its direct effect on R_0, exhibited through demographic PC7, which is in the direction of slowing COVID-19 spread in a population. The other effect is through collinearity with PC1, which reflects a generic correlation of BUAPC with GDPpc, caused by more

construction (higher built-up area) per capita with the increase in GDPpc. Our combination of PC and regression analysis revealed this non-trivial conclusion, which cannot (even qualitatively) be obtained from the pairwise correlation analysis.

Similar reasoning, though perhaps harder to understand intuitively, applies to net migration. Net migration is also significantly (and positively) correlated with HDI, and consequently also with PC1, reflecting a generic tendency of immigrants to flow to countries with higher GDPpc. The direct effect of net immigration, exhibited through PC7 is however harder to intuitively understand, as I-E negatively contribute to R_0, so that faster spread (at least in the initial phase of the epidemic) appears to be associated with a higher number of emigrants. As these are economic migrations (to be distinguished from the movement of refugees), possibly the part of the emigrants returned to their countries with the pandemic's start. In any case, the significant effect of net immigration on R_0 inferred through our analysis is again highly non-trivial, and in the opposite direction from the positive pairwise correlation of R_0 with I-E. For refuges (i.e., percentage of refugee population by country), it exhibits high correlations with only PC3 and PC8, neither of which significantly contribute to R_0. There is also no significant pairwise correlation of refugees with R_0, which robustly shows that this variable does not significantly affect transmissibility.

Regarding pollution, it contributes negatively to demographic PC1 (with the corresponding negative correlation with HDI), while it has a positive contribution to demographic PC4. The pairwise correlation of the pollution with R_0 is negative (-0.31), which is counterintuitive, as it is generally expected that higher pollution should increase COVID-19 transmissibility. This negative correlation with R_0 is however an artifact of the generic negative correlation of the pollution with HDI, while its genuine (direct) effect is reflected through PC4. Our analysis, therefore, revealed the direct effect of long-term air pollution on transmissibility, which is consistent with previously published observations that it can damage the respiratory system and reduce resistance to infections (Domingo and Rovira, 2020; Fattorini and Regoli, 2020), but opposite to naive pairwise correlation analysis.

Raised blood pressure also shows a statistically significant, but counterintuitively negative, correlation with R_0. However, in addition to PC1, raised blood pressure shows a notable correlation only with PC2, which does not significantly affect R_0. This indicates that the negative correlation of this variable with R_0 is a consequence of its generically negative correlation with HDI, instead of a direct effect on COVID-19 transmissibility.

5 Conclusion and Outlook

Numerous studies tried to assess the correlations of different factors with the SARS-CoV-2 virus transmissibility (Li et al., 2020; Notari and Torrieri, 2020; Salom et al., 2021), but the next step should be predicting the environmental risk of the high spreadability in a certain population (Allel et al., 2020; Coccia, 2020; Gupta and Gharehgozli, 2020). Specifically, a relatively small number of the most influential meteorological and demographic factors should be selected for a predictive risk measure that is accurate enough and practical for use. Such risk assessment is very useful in guiding the future strategies of imposing epidemic mitigation measures.

We here demonstrated that taking into account joint effects of different factors can point to qualitatively different conclusions about their influence on the virus transmissibility than considering them individually (as in (Salom et al., 2021)). Utilizing a combination of PCA and feature selection techniques, we were able to disentangle with high confidence which variables independently (and significantly) influence the rate of the infection spread, and which have an only indirect influence or

no influence at all (here found for alcohol consumption, chronic diseases, percentage of the urban population, raised blood pressure and refugees).

While PCA brings clear advantages to regression analysis such as working with a smaller number of variables and abolishing collinearity, the main disadvantage is harder interpretation in terms of original variables. In this case, we were, however, able to unequivocally interpret PCs that significantly affect R0, so that the main driving factors (i.e., PCs) behind COVID-19 transmissibility are the country's wealth/development level corrected by the available indoor space per person and net immigration; pollution levels, and some of the unhealthy living factors; spontaneous behavior change due to developing epidemics; weather seasonality; possibly (marginally) BCG vaccination. These conclusions, and the direction of the corresponding effects, crucially depend on the more complex analysis performed here.

However, when the alignment between certain variables is too high, even the analysis performed here cannot differentiate between the factors genuinely affecting R0 and mere accidental correlations. In such cases, further, specifically designed (such as targeted epidemiological) studies are needed. For example, based on this analysis alone and due to the very high correlation between the cholesterol levels and HDI/GDP it cannot be excluded that cholesterol is a contributing factor to the observed significance of the PC1 component, in addition to the country's prosperity that mimics the contact rate in population (as a crucial disease transmission property). For this reason, our research suggests that a separate study of cholesterol levels in the COVID-19 context (e.g. by measuring cholesterol blood levels along with PCR tests) could be, potentially, of high value since a hypothetical unexpected discovery of inherent cholesterol importance could potentially lead to novel treatments of SARS-CoV-2 infection. Similarly, studies that disentangle the effect of the overall country's prosperity from the intrinsic effects of median age on R0 would be also quite welcome.

Our conclusions about the importance of HDI as a predictor of R0 could be further tested by studies of epidemiological relevance of higher resolution HDI-analogs, such as Subnational HDI (SHDI) or City Development Index (CDI). And if HDI and GDP parameters are confirmed to dominantly influence R0 values simply since they highly and naturally correlate with the frequency of social contacts (as we anticipate to be the case), identifying this as one of the major factors is not without implications. While it is certainly not reasonable to intentionally reduce HDI levels to curb the COVID-19 epidemic, recognizing the importance of this parameter can help us make better predictions of the disease dynamic and locate in advance high-risk spots/areas. The BUAPC variable, which surfaced as another significant factor in our analysis, can have a similar predictive value. As for the PC4 component, reflecting the healthy lifestyle and living conditions, we could and certainly should try to influence the underlying variables - by attempting to reduce obesity, smoking prevalence, physical inactivity, and air pollution. All the more so now that our study indicates the corresponding improvements would also be beneficial to combat the COVID-19 pandemic.

6 Conflict of Interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

7 Author Contributions

MarD, IS, MagD and OM conceived the research. The work was supervised by MarD and IS. Code writing and data analysis by MarD and SM, with help of MagD and OM. Figures and tables made by SM with the help of MagD. A literature search by AR. Manuscript written by IS, AR, and MarD, with help of MagD.

8 Data Availability Statement

Data is available through Salom et al., 2021.

9 Funding

This work was partially supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

10 References

Alexopoulos, E.C. (2010). Introduction to multivariate regression analysis. *Hippokratia* 14(Suppl 1):23-28.

Allel, K., Tapia-Muñoz, T., and Morris, W. (2020). Country-level factors associated with the early spread of COVID-19 cases at 5, 10 and 15 days since the onset. *Glob. Public Health* 15(11):1589-1602. doi: 10.1080/17441692.2020.1814835.

Carozzi, F. (2020). Urban density and COVID-19. *Institute for the Study of Labor (IZA)* [Online], 13440. Available at: https://ssrn.com/abstract=3643204.

Casanova, L.M., Jeon, S., Rutala, W.A., Weber, D.J., and Sobsey, M.D. (2010). Effects of air temperature and relative humidity on coronavirus survival on surfaces. *Appl. Environ. Microbiol.* 76(9):2712-2717. doi: 10.1128/AEM.02291-09.

Chan, K.-H., Peiris, J.M., Lam, S., Poon, L., Yuen, K., and Seto, W.H. (2011). The effects of temperature and relative humidity on the viability of the SARS coronavirus. *Adv. Virol.* 2011:734690. doi: 10.1155/2011/734690.

Coccia, M. (2020). An index to quantify environmental risk of exposure to future epidemics of the COVID-19 and similar viral agents: Theory and practice. *Environ. Res.* 191:110155. doi: 10.1016/j.envres.2020.110155.

Djordjevic, M., Djordjevic, M., Ilic, B., Stojku, S., and Salom, I. (2021). Understanding Infection Progression under Strong Control Measures through Universal COVID-19 Growth Signatures. *Global Challenges* 2021:2000101. doi: 10.1002/gch2.202000101.

Domingo, J., and Rovira, J. (2020). Effects of air pollutants on the transmission and severity of respiratory viral infections. *Environ. Res.* 187:109650. doi: 10.1016/j.envres.2020.109650.

Fattorini, D., and Regoli, F. (2020). Role of the chronic air pollution levels in the Covid-19 outbreak risk in Italy. *Environ. Pollut.* 264:114732. doi: 10.1016/j.envpol.2020.114732.

Frost, F., Petersen, H., Tollestrup, K., and Skipper, B. (2007). Influenza and COPD mortality protection as pleiotropic, dose-dependent effects of statins. *Chest* 131(4):1006-1012. doi: 10.1378/chest.06-1997.

Gangemi, S., Billeci, L., and Tonacci, A. (2020). Rich at risk: socio-economic drivers of COVID-19 pandemic spread. *Clin. Mol. Allergy* 18:12. doi: 10.1186/s12948-020-00127-4.

Garland, P., Babbitt, D., Bondarenko, M., Sorichetta, A., Tatem, A.J., and Johnson, O. (2020). The COVID-19 pandemic as experienced by the individual. arXiv [Preprint]. Available at: https://ui.adsabs.harvard.edu/abs/2020arXiv200501167G/abstract (Accessed March 12, 2021).

Gupta, A., and Gharehgozli, A. (2020). Developing a Machine Learning Framework to Determine the Spread of COVID-19. SSRN [Preprint]. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3635211 (Accessed March 12, 2021).

607 Haddad, C., Bou Malhab, S., Sacre, H., and Salameh, P. (2021). Smoking and COVID-19: A Scoping
608 Review. *Tob. Use Insights* 14:1179173X21994612. doi: 10.1177/1179173X21994612.

609 Hassan, M., Bhuiyan, M., Tareq, F., Bodrud-Doza, M., Tanu, S., and Rabbani, K. (2021). Relationship
610 between COVID-19 infection rates and air pollution, geo-meteorological, and social
611 parameters. *Environ. Monit. Assess.* 193(1):29. doi: 10.1007/s10661-020-08810-4.

612 Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining,*
613 *Inference, and Prediction.* New York: Springer.

614 Heidari-Beni, M., and Kelishadi, R. (2020). Reciprocal impacts of obesity and coronavirus disease
615 2019. *J. Res. Med. Sci.* 25(1):110. doi: 10.4103/jrms.JRMS_416_20.

616 Heilingloh, C.S., Aufderhorst, U.W., Schipper, L., Dittmer, U., Witzke, O., Yang, D., et al. (2020).
617 Susceptibility of SARS-CoV-2 to UV irradiation. *Am. J. Infect. Control* 48(10):1273-1275. doi:
618 10.1016/j.ajic.2020.07.031.

619 Jolliffe, I.T. (2002). *Principal Component Analysis.* New York: Springer.

620 Kapitsinis, N. (2020). The underlying factors of the COVID-19 spatially uneven spread. Initial
621 evidence from regions in nine EU countries. *Regional Science Policy & Practice* 12(6):1027-
622 1045. doi: 10.1111/rsp3.12340.

623 Khajanchi, S., Sarkar, K., Mondal, J., and Perc, M. (2020). Dynamics of the COVID-19 pandemic in
624 India. arXiv [Preprint]. Available at: https://arxiv.org/abs/2005.06286 (Accessed March 12,
625 2021).

626 Li, M., Zhang, Z., Cao, W., Liu, Y., Du, B., Chen, C., et al. (2020). Identifying novel factors associated
627 with COVID-19 transmission and fatality using the machine learning approach. *Sci. Total*
628 *Environ.* 764:142810. doi: 10.1016/j.scitotenv.2020.142810.

629 Lin, S., Wei, D., Sun, Y., Chen, K., Yang, L., Liu, B., et al. (2020). Region-specific air pollutants and
630 meteorological parameters influence COVID-19: A study from mainland China. *Ecotoxicol.*
631 *Environ. Saf.* 204:111035. doi: 10.1016/j.ecoenv.2020.111035.

632 Mallapaty, S. (2021). *What's the risk of dying from a fast-spreading COVID-19 variant?* [Online].
633 Available at: https://www.nature.com/articles/d41586-021-00299-2 [Accessed March 12,
634 2021].

635 Notari, A. (2021). Temperature dependence of COVID-19 transmission. *Sci. Total Environ.*
636 763:144390. doi: 10.1016/j.scitotenv.2020.144390.

637 Notari, A., and Torrieri, G. (2020). COVID-19 transmission risk factors. medRxiv [Preprint]. Available
638 at: https://www.medrxiv.org/content/10.1101/2020.05.08.20095083v1 (Accessed March 12,
639 2021).

640 O'Neill, L.A., and Netea, M.G. (2020). BCG-induced trained immunity: can it offer protection against
641 COVID-19? *Nat. Rev. Immunol.* 20(6):335-337. doi: 10.1038/s41577-020-0337-y.

642 Pawelec, G., and Larbi, A. (2008). Immunity and ageing in man: Annual Review 2006/2007. *Exp.*
643 *Gerontol.* 43(1):34-38. doi: 10.1016/j.exger.2007.09.009.

644 Pope, P., and Webster, J. (1972). The use of an F-statistic in stepwise regression procedures.
645 *Technometrics* 14(2):327-340. doi: 10.1080/00401706.1972.10488919.

646 Ran, J., Zhao, S., Han, L., Qiu, Y., Cao, P., Yang, Z., et al. (2020). Effects of particulate matter
647 exposure on the transmissibility and case fatality rate of COVID-19: A Nationwide Ecological
648 Study in China. *J. Travel Med.* 27(6):taaa133. doi: 10.1093/jtm/taaa133.

649 Rocklöv, J., and Sjödin, H. (2020). High population densities catalyse the spread of COVID-19. *J.*
650 *Travel Med.* 27(3):taaa038. doi: 10.1093/jtm/taaa038.

651 Ruiz, J.I., Nuhu, K., McDaniel, J.T., Popoff, F., Izcovich, A., and Criniti, J.M. (2015). Inequality as a
652 powerful predictor of infant and maternal mortality around the world. *PLoS One*
653 10(10):e0140796. doi: 10.1371/journal.pone.0140796.

654 Sagripanti, J.L., and Lytle, C.D. (2020). Estimated Inactivation of Coronaviruses by Solar Radiation
655 With Special Reference to COVID-19. *Photochem. Photobiol.* 96(4):731-737. doi:
656 10.1111/php.13293.

657 Salom, I., Rodic, A., Milicevic, O., Zigic, D., Djordjevic, M., and Djordjevic, M. (2021). Effects of
658 Demographic and Weather Parameters on COVID-19 Basic Reproduction Number. *Front.*
659 *Ecol. Evol.* 8(524):617841. doi: 10.3389/fevo.2020.617841.

660 Schmidt, N., Wing, P., McKeating, J., and Maini, M. (2020). Cholesterol-modifying drugs in COVID-
661 19. *Oxf. Open Immunol.* 1(1):iqaa001. doi: 10.1093/oxfimm/iqaa001.

662 Shahid, Z., Kalayanamitra, R., McClafferty, B., Kepko, D., Ramgobin, D., Patel, R., et al. (2020).
663 COVID-19 and Older Adults: What We Know. *J. Am. Geriatr. Soc.* 68(5):926-929. doi:
664 10.1111/jgs.16472.

665 Srivastava, A. (2021). COVID-19 and air pollution and meteorology-an intricate relationship: A
666 review. *Chemosphere* 263:128297. doi: 10.1016/j.chemosphere.2020.128297.

667 Tegally, H., Wilkinson, E., Giovanetti, M., Iranzadeh, A., Fonseca, V., Giandhari, J., et al. (2020).
668 Emergence and rapid spread of a new severe acute respiratory syndrome-related coronavirus 2
669 (SARS-CoV-2) lineage with multiple spike mutations in South Africa. medRxiv [Preprint].
670 Available at: https://www.medrxiv.org/content/10.1101/2020.12.21.20248640v1 (Accessed
671 March 12, 2021).

672 Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. Roy. Stat. Soc. B Met.*
673 58(1):267-288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

674 van Doremalen, N., Bushmaker, T., Morris, D.H., Holbrook, M.G., Gamble, A., Williamson, B.N., et
675 al. (2020). Aerosol and Surface Stability of SARS-CoV-2 as Compared with SARS-CoV-1. *N.*
676 *Engl. J. Med.* 382(16):1564-1567. doi: 10.1056/NEJMc2004973.

677 World Bank (2020a). *GDP per capita (current US$)* [Online]. Available at:
678 https://data.worldbank.org/indicator/NY.GDP.PCAP.CD [Accessed January, 2021].

679 World Bank (2020b). *World Bank Open Data* [Online]. Available at: https://www.worldbank.org/
680 [Accessed May, 2020].

681 Xie, Z., Qin, Y., Li, Y., Shen, W., Zheng, Z., and Liu, S. (2020). Spatial and temporal differentiation
682 of COVID-19 epidemic spread in mainland China and its influencing factors. *Sci. Total*
683 *Environ.* 744:140929. doi: 10.1016/j.scitotenv.2020.140929.

684 Zheng, Z., Peng, F., Xu, B., Zhao, J., Liu, H., Peng, J., et al. (2020). Risk factors of critical & mortal
685 COVID-19 cases: A systematic literature review and meta-analysis. *J. Infect.* 81(2):e16-e25.
686 doi: 10.1016/j.jinf.2020.04.021.

687 Zou, H., and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Stat.*
688 *Soc. B* 67(2):301-320. doi: 10.1111/j.1467-9868.2005.00503.x.

689