Population Review

Volume 60, Number 2, 2021

Type: Article, pp. 66-87

The Role of Sociodemographic Factors During a Pandemic Outbreak: Aggravators and Mitigators

Authors: Masoud Shahmanzari and Enes Eryarsoy

Affiliations: Department of Management Information Systems, Faculty of Business, Ozyegin University (Shahmanzari); Sabanci Business School (Eryarsoy)

Corresponding author/address: Enes Eryarsoy. Sabanci Business School, Orta Mahalle, Üniversite Caddesi No:27 Tuzla, 34956 İstanbul; email: enes@sabanciuniv.edu

Abstract

Many macro- and micro-level factors affect the spread of an infectious disease. Among them are sociodemographic, socioeconomic, sociocultural, health care system infrastructure, use of alcohol or substances, level of life disruptions because of chronic illnesses. Because of accuracy and timeliness issues, officials are often forced to make one-size-fits-all decisions across all regions. This paper offers a framework to analyze and quantify the interrelationships between a wide set of sociodemographic factors and the transmission speed of the pandemic to facilitate custom-fitted regional containment measures. The purpose of this paper is to investigate the role of a comprehensive set of sociodemographic factors in the diffusion of COVID-19 analytically. Our findings suggest that diverse sets of sociodemographic factors drive the transmission during different stages of the pandemic. In specific, we show that variables such as gender, age groups, daily commuting distances, modes of employment, poverty and transportation means are found to be statistically significant in the transmission speed of COVID-19. Our results do not suggest a statistically significant relationship between transmission speed and migration-related variables. We also find that the importance levels for the statistically significant variables vary across different stages of the pandemic. Our results point out a variety of public policy insights and implications.

Keywords

COVID-19, public health, demographic features, migrants, SIR, pandemic, United States

1. Introduction

COVID-19 first surfaced in late 2019 in the city of Wuhan, China. Since then, the virus has spread worldwide, infecting more than 188 million individuals and resulting in about 4 million deaths (as of 16 July 2021, from John Hopkins Coronavirus Research Center). The pandemic has attracted broad attention from researchers. Compared with other outbreaks, such as MERS, SARS and AIDS, the impact of COVID-19 on countries is significantly more severe. According to health experts and economists, the pandemic will continue unless at least 60% of the global population is immunized (Atkeson & Hall, 2020) – via natural immunity or vaccines. Until then, governments will keep enacting various restrictive policies (such as social distancing, curfews, lockdowns and national quarantines) at different levels to control the pandemic. Among the most critical factors that aggravate the spread of COVID-19, the following are well-researched: (i) environmental, (ii) hygiene and isolation, (iii) human mobility and (iv) socioeconomic and sociodemographic.

Environmental factors: The extant literature on the spread of COVID-19 suggests that environmental factors (such as temperature, humidity, air quality) and the spread or the severity of the disease are linked. Studies report conflicting results about transmission rates and temperature. For example, Shahzad et al. (2020) reports a positive relationship with temperature for Wuhan, but negative relationships for Zhejiang and Shandong. Poirier et al. (2020) and Luo et al. (2020), on the other hand, report that environmental variables cannot account for variabilities in transmission rates. Kifer et al. (2020) studied the effect of temperature and humidity on the severity of the disease on 6,911 patients in Europe and China. They detected significant decreases in terms of both severity and hospital stay durations as temperatures rise. Low relative humidity levels of air also significantly increase the severity of COVID-19 (Kifer et al., 2020; Ma et al., 2020; Wang et al., 2020). In their study on air quality and transmissibility of COVID-19, Manoj et al. (2020) reported that increased levels of aqueous atmospheric aerosols, or air pollution, can facilitate a pathway for a higher rate of transmission. Zhang et al. (2020) reported similar results.

Hygiene and isolation factors: Studies focusing on hygiene and isolation factors and the transmission speed of COVID-19 transmission reported important, yet not surprising, results. The studies suggested a strong negative link between transmissibility of COVID-19 and hand hygiene (Rundle et al., 2020; G. Q. Zhang et al., 2020), nail hygiene (Wu & Lipner, 2020), oral hygiene (González-Olmo et al., 2020), wearing facemasks (Liu & Zhang, 2020; Mittal et al., 2020), and isolation rooms (G. Q. Zhang et al., 2020).

Mobility factors: Similar to many other pandemics, decreasing population mobility can help curb the spread of COVID-19. Delen et al. (2020) investigated the relationship between the cellular network mobility patterns and the transmission rates across many countries. They found a strong link between increased mobilities and higher transmission rates. Other studies reported identical results (Badr et al., 2020; Kraemer et al., 2020; Xiong et al., 2020).

Socioeconomic and sociodemographic factors: This set of factors are central to the present research. We, therefore, provide a more in-depth review of the literature on socioeconomic

and demographic factors and the COVID-19 pandemic. Demographic data include statistical presentation of the socioeconomic information, such as income, unemployment and ethnicity. There are few studies that explore the role of (some) demographic variables in the diffusion of COVID-19. The literature suggests that variations in demographic structures across different populations account for the differences in transmission speeds. This subset of studies considers micro-, mezzo- and macro-level factors, such as poverty, ethnicity, immigration, genetic profiles, environmental hazards, health demographics and household population. For example, Martin et al. (2020), suggested that different sociodemographic heterogeneity indicators, such as larger households, produce different transmission rates across populations.

Dowd et al. (2020) examined the role of demographics (age structure) in Italy and South Korea to understand the fatality rates of COVID-19. Their results showed an extremely higher burden of mortality in older populations compared with younger ones. Mogi and Spijker (Mogi & Spijker, 2020) used the COVID-19 dataset and European Social Survey in 20 European countries to study the relationship between social, demographic, and economic features and reported case numbers in one-month period. The variables comprise three groups: “socially and economically vibrant”, “high-educated and not aged”, and “densely populated and traditional”. They claimed that the “socially and economically vibrant” group is significantly associated with increased COVID-19 cases. Bayer and Kuhn (2020) explore the intergenerational ties and COVID-19 fatality rates. In particular, they analyze how variations in living arrangements of multiple generations contribute to cross-country variations. Their findings show a significant relationship between the share of working-age families living with their parents and COVID-19 incidences in each country. Other studies investigate gender effects and found that male sex was an independent risk factor (Walter et al., 2020) . Eryarsoy et al. (2020) studied age disparities of COVID-19 patients regarding incidents such as hospitalizations, ICU usages, and deaths.

Besides age and gender, disparities involving other demographic items have also been extensively studied. Migration and ethnicity are among the more widely studied demographic factors. The relationship between the presence of national and international migrants and the transmission rates of infectious diseases has been extensively studied. Earlier literature emphasizes the role of immigrants and refugees in various outbreaks, including biblical plagues, 1918 influenza, AIDS outbreak, and SARS pandemic (Freifeld et al., 2020). In a few cases, a specific subgroup of immigrants introduced a contagious disease into native populations. For instance, a significant portion of HIV virus diffusion can be attributed to immigrant sex workers (Steen et al., 2020). Migrants, as one of the mobile social groups due to their strong bonds with their country of origin, may magnify the risk of introducing potential infectious diseases to host countries. In societies, non-citizens are typically more mobile than citizens. Therefore, for countries with a large percentage of immigrants like the US (14.4% of the population are immigrants), it is crucial to investigate the interrelationships between the characteristics of immigrant groups and the dynamics of

the spread of disease during pandemics. Keller and Wagner (2020) discussed the policies to keep the rights of immigrants in the US. They argue that keeping immigrants in overcrowded facilities will ultimately increase the chance of an outbreak. Lodge and Kuchukhidze (2020) compared the response to HIV and COVID-19 infections among migrant workers and identified them as one of the most vulnerable groups. Migrant workers are compelled to expose the risks of the exodus to their home, since they don't have enough social and economic resources to protect themselves.

The literature also includes studies that focus on the role of other important demographic items such as heterogeneity, education, income and poverty. Merler and Ajelli (2010) use a stochastic spatially individual-based model to study the effect of population heterogeneity and various human mobility factors on the spread of influenza pandemic in 37 European countries. According to their findings, the attack rates depend significantly on sociodemographic characteristics, including the size of household groups and the proportions of students in target populations, as they typically have higher contact rates. Bogg and Milad (2020) studied the patterns and psychosocial relations of COVID-19 guideline adherence. They tested the relationship among demographic features (age, sex, ethnicity, education, income, relationship status, and presence of children), personality characteristics, adherence to COVID-19 guidelines and related social cognition in a sample of 501 US citizens over a two-week period. Their findings showed the health relevance of personality features. Pullano (2020) considered age groups and mobilities. While our study has similarities, we consider a wider selection of demographic variables and transmission speeds, and two analysis stages.

In this paper, we investigate the role of population demographics and the transmission speed of COVID-19. In particular, we try to find statistically significant demographic characteristics that affect the spread of COVID-19. We itemize our principal contributions to the existing literature as follows:

> *(i) Comprehensive set of sociodemographic variables:* Existing literature considers only very limited (1-5) demographic items. We use a comprehensive set of demographic items to study their links to COVID-19 transmission rates.
>
> *(ii) Low level of data granularity:* We conduct county-level analysis (for the US). For policymakers, it is essential to know which demographic factors are driving the spread of disease and whether factors such as the presence of migrants in an area or heterogeneity increase the diffusion of COVID-19. This enables policymakers to better understand underlying differences across regions and make region-specific decisions. We give details about our datasets in Section 2.1.
>
> *(iii) Two-stage research design:* We design a two-stage research setup based on the susceptible, infected, recovered (SIR) and regression models. We use SIR model's

output as the dependent variable of the regression analysis. The research design is provided in Section 2.2.

2. Data and methods

2.1. Data

After the COVID-19 outbreak, the *New York Times* published daily statistics of COVID-19 pandemic in the US at state- and county-levels in their GitHub account. The US Census Bureau releases data containing a variety of statistics, including migration and ethnicity. We merged the two datasets to explore the relationship between the sociodemographic information of US counties and the transmission spread of COVID-19. At the time of writing most of this article (in 2020), confirmed cases of the disease were still increasing in the US and in most of the regions in the world. In the meantime, many regions started reducing restrictive policies and permitted many businesses to get back to work. We summarize the attributes for both datasets in Table 1.

COVID-19 data

We used the data provided by the *New York Times* for COVID-19 cases and deaths at the county level *(New York Times*, 2020). The dataset incorporates three attributes: 1) Time (since January 21, 2020), 2) Geographic Information (county, state, Federal Information Processing Standard (FIPS)), and 3) COVID-19 incident statistics (reported cases and deaths). The dataset provides reported incidences about the pandemic since earlier days of

Table 1. *New York Times* COVID-19 data and ACS data

	COVID-19 *New York Times*	ACS
Data		
Starting Date	January 21, 2020	N/A[a]
Num of Weeks	17	N/A
States (n)	50	50
Counties (m)	1744	3007
Variables		
1	Date	Population
2	County	Gender
3	State	Age
4	FIPS	Economic
5	Cases	Ethnicity
6	Deaths	Workforce
7	–	Mobility Means

[a] Not Applicable

its recognition. Therefore, our COVID-19 data started on January 21, 2020, and spans 17 weeks. It consists of 273,314 rows. The dataset includes only 1,805 counties, out of 3,143. This is mostly because there was no available data on the number of COVID-19 cases in other counties.

American Community Survey Data

In this research, our goal is to test the relationship between reported COVID-19 cases and the sociodemographic characteristics of US citizens and migrants. To this end, updated demographics of the population in the US, both at state and county levels, are used.

We extract the required data from the 2014–2018 American Community Survey (ACS) dataset (2019). Our dataset includes five groups of attributes: socio-demographics (total population, number of foreign-born, citizens, and non-citizens, gender [men, women], age groups [18 to 24, 25 and above]), economic factors (employed, unemployment rate, income, income per cap, poverty, child poverty), ethnicity and diversity (Hispanic, white, black, native, Asian-Pacific), workforce (professional, service, office, construction, production, work at home, private work, public work, self-employed, and family work), and mobility means (drive, carpool, transit, walk).

It is worth noting that the terms 'foreign-born' and 'non-citizen' have different meanings in ACS dataset. The foreign-born population consists of individuals who did not have US citizenship at birth, including lawful immigrants (naturalized citizens), temporary lawful migrants (e.g. international students), lawful permanent residents (refugees and asylum seekers), and unauthorized immigrants. On the other hand, a non-citizen is an individual who does not hold US citizenship. Hence, the foreign-born population includes non-citizens as well.

We merge COVID-19 and ACS datasets, and limit our analysis to the counties present in the *New York Times* dataset. After row-wise removal of counties with missing values and inclusion of state-level data from 50 states, we end up with 1,843 rows. We then further filter our dataset to include only the counties with over 500 reported cases during earlier days (between week 4 and week 11) of the pandemic. The threshold of reported cases is set to 500 to reduce sampling bias. Our final dataset consists of 555 counties.

2.2 Methods

The first stage of our two-stage analysis is to measure the transmission speed of COVID-19. Looking at reported incident statistics is misleading, as the same number of incidents during the early days versus during the peak days of a pandemic translates into different diffusion speeds. Therefore, a special coefficient corresponding to the spread speed is needed. This, in epidemic modeling, corresponds to a coefficient known as the transmission rate. We use susceptible, infected, recovered (SIR) model to calculate the transmission rate

(Kermack and McKendrick, 1991). We then use transmission rates as our dependent variable in our regression analysis.

SIR model

In our study, we are interested in a measure of the transmission speed of COVID-19. During pandemics, various models are used to forecast the spread. One of the most common ones are the compartmental models such as SIR model. For each time period, it is used to forecast the number of people susceptible to the disease, already infected by the disease, and have recovered from the disease. A distinctive feature of the SIR model is its simplicity that enables researchers to model infection behavior with only two parameters. We use SIR model to determine the response variable representing the transmission rates of the infection for each region and stage. We use these rates in the second stage of our analysis to identify the relationships between regional sociodemographic factors and regional spread speeds of the virus.

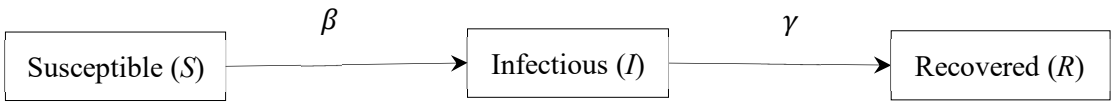

Figure 1. General framework of SIR model

We depict the general framework of the SIR model in Figure 1. The SIR model consists of three distinct compartments: susceptible (*S*), infected (*I*), and recovered (*R*). In its original

$$\frac{dS}{dt} = -\beta \frac{S \times I}{N} \tag{1}$$

$$\frac{dI}{dt} = \beta \frac{S \times I}{N} - \gamma I \tag{2}$$

$$\frac{dR}{dt} = \gamma I \tag{3}$$

form, it includes two parameters: effective contact rate (β) and recovery rate (γ). While the former parameter controls the transition from *S* to *I*, the latter controls the transition from *I* to *R*. In particular, β corresponds to the average number of transmissions per person, and it accounts for the number of new cases over time. β is sensitive to various governmental interventions such as curfews, closing schools, social distancing and quarantines. While

these intervention strategies can change the transition of individuals from S to I, the transition from I to R exclusively depends on the time span that a person is contagious. This rate is captured by γ. So, a shorter mean infectious time ($1/\gamma$) infers a faster transition from I to R. Equations (1) to (3) are differential equations that represent the transitions among these three compartments where N (S+I+R) and t stand for the population size and time, respectively.

Table 2. List of attributes for regression analysis

Field	Explanation	Variable Type
Region	County or State	ID
FromPeriod	Starting period for the calculated Beta variable. (4-16)	Date
TillPeriod	Ending period for the calculated Beta variable (5-17)	Date
TotPop*	Total population of the county	
Foreign*	% of foreign-born population	
Citizen*	% of US citizens	
NonCitizen*	% of US non-citizens	
Gender*	Male/Female ratio	
18_25*	% of population aged between 18 and 25	
25Plus*	% of population older than 25	
Hispanic*	% of Hispanic/Latino population	Ethnicity, Diversity, Demographics
White*	% of White population	
Black*	% of Black population	
Native*	% of Native population	
Asian*	% of Asian population	
Pacific*	% of Pacific population	
Diversity*	A single value corresponding to ethnical diversity, calculated using entropy	
Income*	Median household income ($)	
IncomePerCap*	Income per capita ($)	
Poverty*	% of population under poverty level	
ChildPoverty*	% of children under poverty level	Economic
EmployedPerc*	1- % of unemployment	
SelfEmployed*	% of population self-employed	
FamilyWork*	% of population in unpaid family work	
Professional*	% of employed people in management, business, science, and arts	
Service*	% of people employed in service jobs	
Office*	% of people employed in sales and office jobs	
Construction*	% of people employed in natural resources, construction, and maintenance	Workforce
Production*	% of people employed in production, transportation, and material movement	
PrivateWork*	% of population employed in private industry	
PublicWork*	% of population employed in public jobs	
Drive*	% of people commuting alone in a car, van, or truck	
Carpool*	% of people carpooling in a car, van, or truck	
Transit*	% of people commuting on public transportation	
Walk*	% of people walking to work	Mobility Means
OtherTransp*	% of people commuting via other means such as biking	
WorkAtHome*	% of population working at home	
MeanCommute*	Mean commute time (minutes)	
$\beta_{t,t+k}$	Effective contact rate/spread speed for time interval $[t, t+k]$	Target variable

*County level ACS community survey data (ACS, 2019)

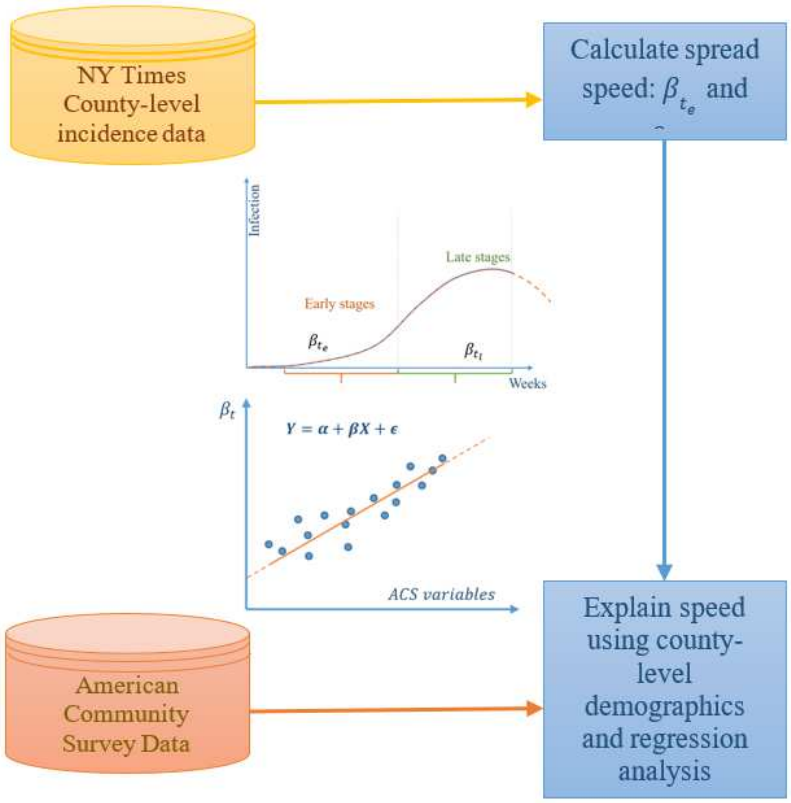

Figure 2. Data analysis procedure

Regression and SIR model

At this stage, our dataset included a large number of covariates. All-inclusive regression models may suffer from multicollinearity problems, therefore, may not generalize well. We refer to this as over-specification bias. Even statistically significant models may not be interpretable (statistical significance issues).

There are different ways to avoid such pitfalls. For example, one may perform a method called best-subset selection, which is an exhaustive search using all possible variable combinations to find the best fit with statistical significance while avoiding multicollinearity. The best-subset problem involves evaluating exponentially many regressions and is known to be NP-hard (Natarajan, 1995). The problem is not computationally feasible for a large number of covariates.

An alternative is to use stepwise regression procedures for achieving results much faster. The stepwise procedure allows features to enter or leave sequentially according to criteria such as Bayesian Information Criterion (BIC) (Schwarz, 1978), Akaiki Information Criterion (AIC) (Akaike, 1998), or Adjusted R2 (Burnham & Anderson, 2002).

Multicollinearity is typically addressed by a post-regression estimation of variance inflation factor ($VIF = 1/(1 - R^2)$). As a rule of thumb, a variable with a VIF of less than [5-10] is included in the model (Menard, 2002; Neter et al., 1989). While stepwise regression has been extensively used in the literature, it is only "locally optimal", meaning for larger models the approach doesn't warrant the best-subset selection. A recent approach by Zambom and Kim (Zambom & Kim, 2018) offered a significance-controlled stepwise variable selection procedure. The method aims at maximizing adjusted R^2 while maintaining a desired level of significance. Another alternative is using regularization parameters. Tibshirani (1996) suggested adding regularization parameters to regression to tackle the multicollinearity issue. The two major regularization methods are Ridge and Lasso. Lasso regression has a tendency to choose only one of the highly correlated variables to tackle the collinearity problem, therefore, it performs feature selection.

We divided the spread into two periods to capture the relationship between the spread speed and other ACS data variables: (*i*) initial period corresponding to weeks 6 through 11, and (*ii*) subsequent period corresponding to weeks 12 through 17. The initial period in our analysis corresponds to a period of fewer precautions in COVID-19 in the US timeline, compared to the late period.

We explore the relationship between β_t and other variables, including the presence of immigrants. To this end, we use a linear regression analysis with β_t as dependent variable. To estimate β_t values we fit SIR model by using Subplex (Nelder-Mead method on the order of subspaces) (Rowan, 1990), Broyden–Fletcher–Goldfarb–Shanno a quasi-Newton algorithm (BFGS)(Byrd et al., 1995), and Multilevel Single Linkage (Kucherenko, 2005) methods simultaneously to ensure the stability of calculated error-minimizing β parameters. In this paper, we only use the best-obtained value with respect to the average squared error. Except for a few instances (<10, depending on time periods), all algorithms return almost identical β_t values. We exclude such instances with significantly different β_t values from our analysis.

Prior to running linear regression, normalization is usually performed to overcome the normality assumption of the outcome. In addition, normalization scales variables and increases the interpretability of the regression coefficients while setting the intercept at zero. We normalize our data by trying different normalizing transformations including Yeo-Johnson (Yeo & Johnson, 2000), Box Cox (Box & Cox, 1964), and ordered quantile transformations (Peterson & Cavanaugh, 2020). The one that yields the lowest Pearson P-test statistic for normality is selected and the variable is normalized accordingly. For regression, we use Zambom and Kim's (2018) stepwise method and checked for VIF. Our earlier research on several other datasets suggested using this method produces identical to other novel methods, such as the Bertsimas et al. (2016) optimal best-subset selection and is magnitudes faster.

3. Results

We coded all our analyses in R. In the first stage of our analysis, we look at the relationship between the percentage of the non-citizen population in each region and the diffusion speed. The literature includes studies that link immigrants (as disadvantaged minority groups) and their likelihood of being more adversely affected by COVID-19. While the immigrants have been targeted by remarks regarding their roles in spreading the COVID-19 faster, to the best of our knowledge this study is the first to analyze the relationship between immigrants (or diversity) and speedier transmission rates.

3.1. Immigrants and the spread of COVID-19

For this part of our analysis, we investigate the relationship between spread speeds (β_t) and immigrant ratios. We first create three different cross-sectional datasets (corresponding to initial, subsequent, and all periods). We then calculate correlation scores between ratio of immigrants and regional transmission speeds of COVID-19 (β_t values).

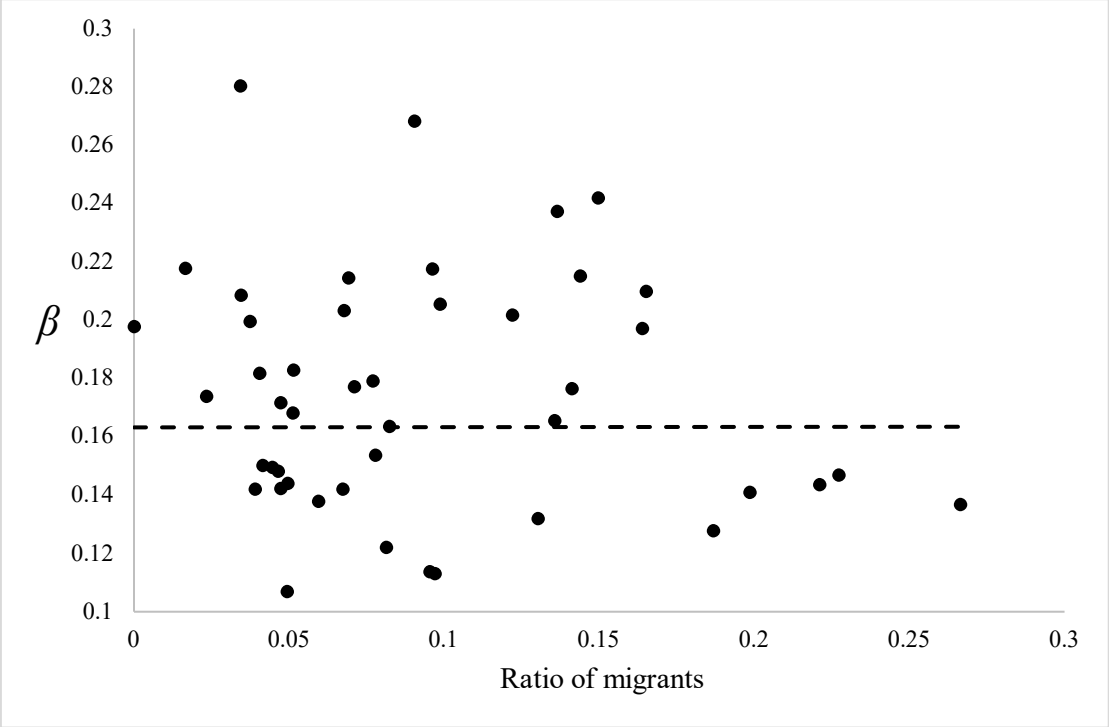

Figure 3. Results of regression analysis from week 4 to week 10 (US states)

We aggregate these cross-sectional datasets and calculate correlation scores between ratio of immigrants and the transmission rates of COVID-19 (β_t values). We first calculate R^2 value at the state-level corresponding to the transmission rates and immigration as 0.0002. We then calculate R^2 value at county level as $R^2 = 0.000004$. These R^2 values show that the pair-wise relationships between migration (non-citizen ratio), diversity and spread speed is very small. Correlation table in the appendix gives the coefficients corresponding to the entire period of study, from week 4 to week 17 of the pandemic.

We give an illustration of this in Figure 3 and

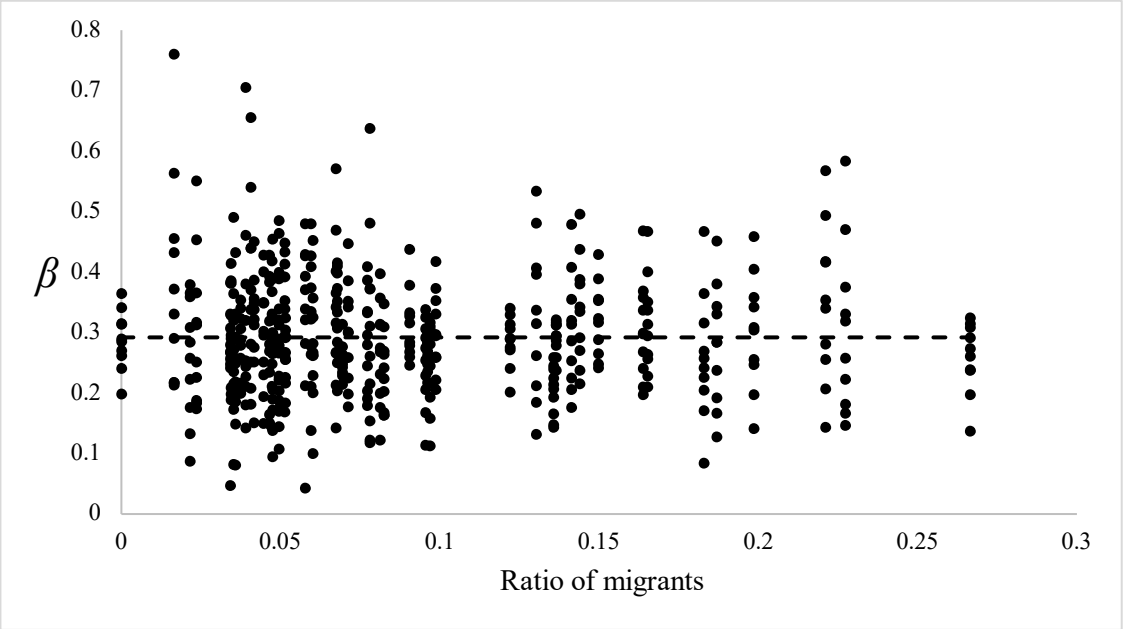

Figure 4. These results suggest that the presence of migrants has no effect on the diffusion speed of COVID-19 both at state and county levels. Figure 3 reports the scatter plot for

weeks 4-10 at state level.

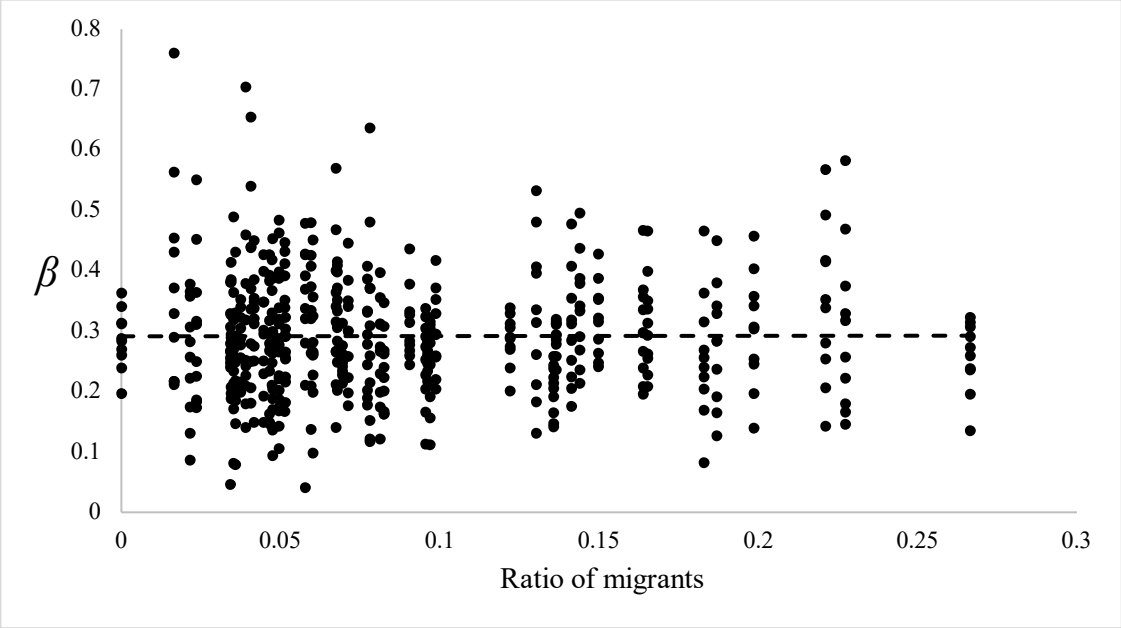

Figure 4 represents the scatter plot of the same time period at county level.

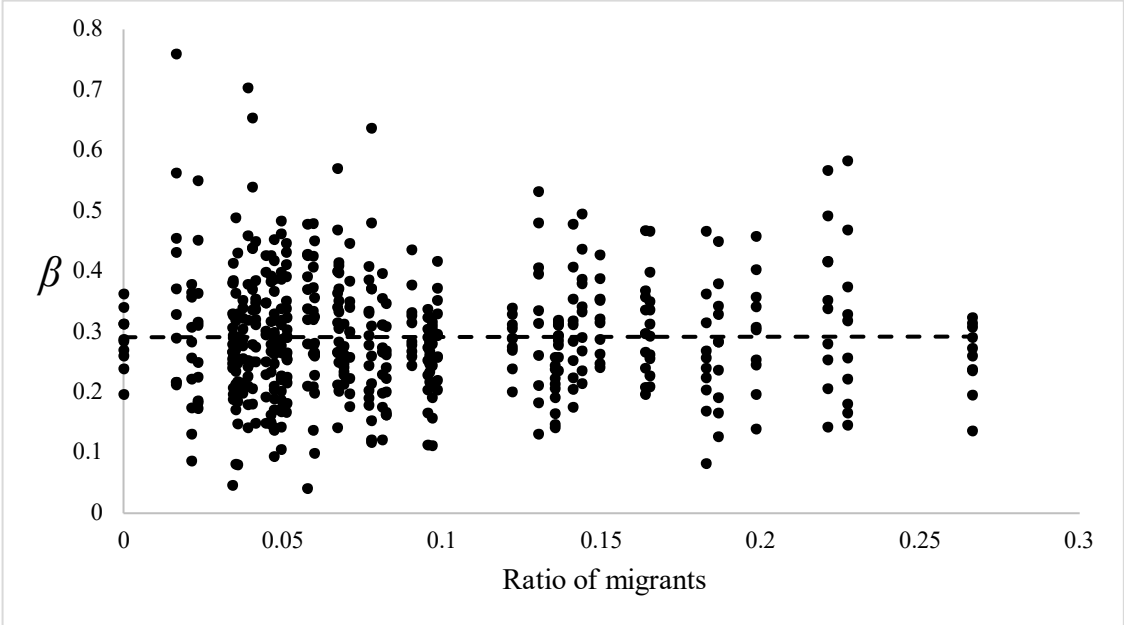

Figure 4. Results of regression analysis from week 4 to week 10 (US counties)

3.2. Comparison of early and late spread speed

The county-level dataset suffers from slight inconsistencies. For example, we expect the cumulative number of cases by date to be a non-decreasing series of numbers. In our dataset, however, occasionally some counties reported decreasing cumulative number of cases on some cases. This is possibly because of reporting differences, and can interfere with computing pandemic transmission rate accurately. Therefore, we conclude that the earliest days of pandemic bring about a lot of uncertainties. Thus, we decide to limit our analysis to week 4 and onwards. We then regressed ACS covariates on β_{4-17} values we calculated using SIR model (Table 4). As there are factors other than sociodemographic ones at play, expecting high R2 values is not realistic. Even though the model was significant, adjusted R^2 was not high (0.07). This is possibly due to different spread dynamics in place during different periods of the spread. We, therefore, create three separate datasets: cross-sectional dataset for the initial (weeks 4-10), for the subsequent (weeks 11-17), and overall (weeks 4-17) periods. We standardize each of the dataset variables. Stepwise regression procedures may eliminate variables of particular interest such as immigration and diversity, despite being statistically significant (because of collinearity). We, therefore, first check if average β_t values of counties lying on opposite sides regarding migration ratios and county diversities are different (Table 3).

Table 3. β-values corresponding to different county groupings regarding migration and diversity

County *(# of counties)*	$\boldsymbol{\beta_e}$, t={4,10}	$\boldsymbol{\beta_l}$, t={11,17}	$\boldsymbol{\beta}$, t={4,17}
Counties with higher migration *(278)*	0.079	0.028	0.122
Counties with lower migration *(277)*	-0.025	-0.121	-0.070
Counties with higher diversity *(277)*	-0.047	-0.077	-0.014
Counties with lower diversity *(278)*	0.093	0.016	0.059
All Counties *(555)*	**<0.0001**	**<0.0001**	**<0.0001**

Source: Based on ACS community survey data (ACS, 2019)

We perform unpaired *t*-tests to check if average β_t values corresponding to high/low migration counties, and high/low diversity countries are statistically different. The tests indicate that differences in β_t values are nonsignificant at 0.05 level (p=0.22 for migration during early periods, p=0.085 for migration during late periods, p=0.10 for diversity during early periods, p=0.47 for late periods). We looked at the spread of COVID-19 regarding other covariates using stepwise regression analysis in Zambom and Kim (2018). We give the two different regression results in Table 4.

Table 4 refers to three time intervals: initial (weeks 4-10), subsequent (weeks 11-17) and all (weeks 4-17). The table contains standardized coefficients from three separate regression models built following significance-controlled covariate selection procedure by Zambom

and Kim (2018). The inclusion of covariates in their respective models, in addition to their coefficient signs and magnitudes, laid the basis for comparing different periods. The first interval corresponding to the initial stage of the spread is linked to the lower levels of awareness and milder intervention strategies, hence expected to have higher normalized β_t values (Table 3). Table 4 suggests that variables that may relate to migration such as ethnicity diversity, percent of foreign-born population, or percent of US citizens are not found to be significant (except for ethnicity diversity coefficient that is negatively signed). This may be due to existing multi-collinearities in the data.

Table 4. Regression results for standardized data

Variable	Corresponding standardized coefficients and VIFs					
	Early Periods 4-10	VIF	Late Periods 11-17	VIF	All Periods	VIF
Ethnicity, Diversity, Demographics						
County's total population					.315***	2.398
% of US non-citizens						
Ethnical diversity						
Male/Female ratio						
% of population aged between 18 and 25			.130**	2.84		
Economic						
Income			-.285***	3.128		
Income per capita ($)						
% of population under poverty level						
% of children under poverty level						
% of population self-employed	-.209***	1.583	.108*	1.7		
% of population in unpaid family work						
% of unemployment					.120**	
Workforce						
% of employed people in professional career	-.303***	2.057				
% of people employed in service jobs						
% of people employed in sales and office jobs						
% of people employed in construction						
% of people employed in production			-.302***	1.94	.318***	
% of population employed in private industry	.135**	1.202				
% of population employed in public jobs					.243***	
Mobility Means						
% of people commuting alone in a car			.234***	2.071		
% of people carpooling in a car			.201***	2.428		
% of people commuting on public transportation	.206**	1.921				
% of people walking to work	.247***	2.447	-.322***	1.918		
% of people commuting via other means	-.222***	1.921	.173***	1.876	-.158**	
% of population working at home			.140*			
Mean commute time	.122*	2.206	-.182***	1.784	-.162***	

Adjusted R-squared:	0.149	0.231	0.076

Signif. codes: *** p<.0001; ** p<.001; * p<.01, *N*=555

To study the covariates more in depth we, therefore, conduct a separate regression analysis including only the following covariates: "percent of foreign-born population", "percent of US non-citizens", "Male/Female ratio", and "ethnicity diversity". Our aim is to see if the omissions of these variables in Table 4 were due to multicollinearity issues. The results for the initial period yielded: (*i*) positive signed "Male/Female ratio" variable significant at p=0.07 level indicating males to female ratio correlates with spread speed, and negative signed "ethnicity diversity" variable significant at p=0.02 level. This was also suggested in a variety of news resources and academic papers mentioning gender disparities. However, none of these coefficients were found to be statistically significant for subsequent period (weeks 11-17). The results also suggest a distinct statistically significant set of variables during the initial period and the subsequent period. The comparison of the regression analysis results for two periods reveals interesting patterns.

According to our findings, while the percentage of people working in private industry is a significant factor in the virus's spread in the initial period, it is not a significant factor in the late period anymore. Instead, the percentage of people working in public jobs becomes significant. In fact, this observation can help explain the major influences of government's interventions across two job sectors.

The percentage of the young population aged between 18 and 25 turns out as a significant factor only in the subsequent period. Controlling this segment of the population is a challenging task for authorities. As COVID-19 precautions continue to deprive the youth of social engagement with their social circles, it may become more difficult to limit the socialization of the youth. Note that the other age-related variable "Percent of population older than 25" is not significant in the later period, although the exposed group to the virus are as likely as 18 to 25 group to become infected.

Another significant factor in the diffusion of COVID-19 in the subsequent period is "Median household income". According to Table 4, as pandemic proceeds, the level of income negatively affects the transmission speed of COVID-19. Indeed, the higher the income people earn, the lower the risk of being infected they have. One of the most thoroughly studied relationships in economics is the link between income and education (Apolloni et al., 2013). People with higher education tend to earn more. Therefore, education is one of the key variables in curbing the spread of the virus.

The coefficient of variable "Percent of population self-employed" in the initial period is negative. However, in the subsequent period, its sign turns positive. This may be because of precautions and regulations coming into effect later on. While in the earlier weeks of the pandemic the higher amount of self-employment in regions prevents the faster spreads, later as all sorts of worker mobility are limited the sign becomes negative.

Similarly, in Table 4 the variable “Percent of people walking to work” in the early and late periods has opposite signs. The variable has the highest impact (largest of the coefficients) among other variables on the diffusion of the infection in earlier days (0.298). During the initial period, the virus spreads faster in regions where the number of people walking to work is higher. However, later the direction is reversed, which indicates the efficiency of government intervention in applying policies such as limiting the mobility of workers, working from home and social distancing.

Another interesting observation is the change in the coefficient sign of the variables “Percent of people commuting via other means” (such as biking) and “Mean commute time” distances after the transition from the initial period. While the percentage of people using other means for transport is negatively related to the spread speed of the disease in the initial period, the coefficient changes sign in the subsequent period. One likely reason for this change could lie in the fact that when the use of public transport is reduced during the pandemic, people switch to other means of transportation, which may aggravate the in the spread. Therefore, policymakers should take the effect of alternative transportation means into account while imposing restrictions. This direction change is more interesting in the variable “Mean commute time”. The coefficients of this variable are 0.136 and -0.145 in the early and late periods, respectively. This indicates the effectiveness of interventions in transportation policies as the higher average time of commuting decreases the spread of the virus.

4. Summary and conclusion

This study demonstrates the importance of sociodemographic features coupled with mobility factors in the spread of COVID-19 in the US. Since the beginning of the pandemic policymakers have been struggling to identify critical aggravating and mitigating factors to enact efficient restrictive policies. The literature suggests that varying diffusion patterns could be region specific. There has also been some speculation about the link between regional immigrant population densities and the spread speed of the COVID-19 pandemic. While the literature is clear on immigrants being among the vulnerable groups, to the best of our knowledge this is the first study that questioned the link between immigrant population density and the spread speed of the pandemic. Our study highlights that combating outbreaks based on regional characteristics is crucial. Our analysis of real COVID-19 data and a wide selection of sociodemographic variables indicate that the ratio of immigrants or foreigners, and the spread speed, are not related at state- or county-levels. The regression model is capable of serving decision-makers to tailor their prevention and control tactics continuously based on county-level socioeconomic data. The findings also confirm the necessity for effective interventions such as social distancing to slow down the spread of COVID-19.

The results of our two-period regression analysis may help policymakers to implement dynamic intervention strategies at different stages of a pandemic. A significant observation

is the direction change of significant coefficients such as “Percent of people walking to work” and “Mean commute time” while transitioning from the initial period to the subsequent period. We hope that our results help researchers and decision makers to take sociodemographic traits into account while implementing prevention strategies regionally.

One limitation of this study is the lack of other variables, such as education and migrant’s age group in the dataset. Similar factors might be associated with the spread of COVID-19 in the US. To analyze the transmission speed of the infection, we used the SIR model with varying transmission rates. One future direction could use other compartmental models (e.g., susceptible exposed infected recovery, susceptible infected susceptible, and susceptible infected recovered deceased, among others) to measure other dimensions of infection diffusion. Furthermore, a similar analysis can be performed in other countries where demographic information of the population is publicly available.

References

ACS. (2019). *American Community Survey (ACS)*. https://www.census.gov/programs-surveys/acs

Akaike, H. (1998). *Information Theory and an Extension of the Maximum Likelihood Principle* (pp. 199–213). https://doi.org/10.1007/978-1-4612-1694-0_15

Apolloni, A., Poletto, C., & Colizza, V. (2013). Age-specific contacts and travel patterns in the spatial spread of 2009 H1N1 influenza pandemic. *BMC Infectious Diseases*, *13*(1). https://doi.org/10.1186/1471-2334-13-176

Atkeson, A., & Hall, B. (2020). *What will be the economic impact of covid-19 in the us? rough estimates of disease scenarios*. https://doi.org/10.1038/s41421-020-0148-0

Badr, H. S., Du, H., Marshall, M., Dong, E., Squire, M. M., & Gardner, L. M. (2020). Association between mobility patterns and COVID-19 transmission in the USA: a mathematical modelling study. *The Lancet Infectious Diseases*, *20*(11), 1247–1254. https://doi.org/10.1016/S1473-3099(20)30553-3

Bayer, C., & Kuhn, M. (2020). *Intergenerational Ties and Case Fatality Rates: A Cross-Country Analysis*. www.iza.org

Bertsimas, D., King, A., Statistics, R. M.-T. annals of, & 2016, U. (2016). Best subset selection via a modern optimization lens. *JSTOR*, 813–852.

Bogg, T., & Milad, E. (2020). *Slowing the Spread of COVID-19: Demographic, personality, and social cognition predictors of guideline adherence in a representative US sample*. https://psyarxiv.com/yc2gq/

Box, G. E. P., & Cox, D. R. (1964). An Analysis of Transformations. *Journal of the Royal Statistical Society: Series B (Methodological)*, *26*(2), 211–243. https://doi.org/10.1111/j.2517-6161.1964.tb00553.x

Burnham, K., & Anderson, D. (2002). A practical information-theoretic approach. In *sutlib2.sut.ac.th*. Springer. http://sutlib2.sut.ac.th/sut_contents/H79182.pdf

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A Limited Memory Algorithm for Bound Constrained Optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208. https://doi.org/10.1137/0916069

Coronavirus Update. (2020). https://www.worldometers.info/coronavirus/

Delen, D., Eryarsoy, E., & Davazdahemami, B. (2020). No Place Like Home: Cross-National Data Analysis of the Efficacy of Social Distancing During the COVID-19 Pandemic. *JMIR Public Health and Surveillance*, *6*(2), e19862. https://doi.org/10.2196/19862

Dowd, J. B., Andriano, L., Brazel, D. M., Rotondi, V., Ding, X., Liu, Y., & Mills, M. C. (2020). Demographic science aids in understanding the spread and fatality rates of COVID-19. *National Acad Sciences*. https://doi.org/10.1073/pnas.2004911117/-/DCSupplemental

Eryarsoy, E., & Delen, D. (2020). Cross-Country Age Disparities in COVID-19 Cases with Hospitalization, ICU Usage, and Morbidity. In *arXiv*.

Freifeld, A., Bow, E., … K. S.-C. I., & 2017, U. (2020). Conflict and emerging infectious diseases. *Academic.Oup.Com*. https://academic.oup.com/cid/article-abstract/64/12/iii/3855681

González-Olmo, M. J., Delgado-Ramos, B., Ruiz-Guillén, A., Romero-Maroto, M., & Carrillo-Díaz, M. (2020). Oral hygiene habits and possible transmission of COVID-19 among cohabitants. *BMC Oral Health*, *20*(1). https://doi.org/10.1186/s12903-020-01274-5

Keller, A. S., & Wagner, B. D. (2020). COVID-19 and immigration detention in the USA: time to act. *Thelancet.Com*. https://doi.org/10.1016/S2468-2667(20)30081-5

Kermack, W. O., & McKendrick, A. G. (1991). Contributions to the mathematical theory of epidemics--I. 1927. *Bulletin of mathematical biology*, *53*(1-2), 33-55.

Kifer, D., Bugada, D., Villar-Garcia, J., Gudelj, I., Menni, C., Sudre, C., Vučković, F., Ugrina, I., Lorini, L. F., Posso, M., Bettinelli, S., Ughi, N., Maloberti, A., Epis, O., Giannattasio, C., Rossetti, C., Kalogjera, L., Peršec, J., Ollivere, L., … Lauc, G. (2020). Effects of environmental factors on severity and mortality of COVID-19. *MedRxiv*, 2020.07.11.20147157. https://doi.org/10.1101/2020.07.11.20147157

Kraemer, M. U. G., Yang, C. H., Gutierrez, B., Wu, C. H., Klein, B., Pigott, D. M., du Plessis, L., Faria, N. R., Li, R., Hanage, W. P., Brownstein, J. S., Layan, M., Vespignani, A., Tian, H., Dye, C., Pybus, O. G., & Scarpino, S. V. (2020). The effect of human mobility and control measures on the COVID-19 epidemic in China. *Science*, *368*(6490), 493–497. https://doi.org/10.1126/science.abb4218

Kucherenko, S. (2005). Application of Deterministic Low-Discrepancy Sequences in Global Optimization. In *Computational Optimization and Applications* (Vol. 30).

Liu, X., & Zhang, S. (2020). COVID-19: Face masks and human-to-human transmission. In *Influenza and other Respiratory Viruses* (Vol. 14, Issue 4, pp. 472–473). Blackwell Publishing Ltd. https://doi.org/10.1111/irv.12740

Lodge, W., & Kuchukhidze, S. (2020). COVID-19, HIV, and Migrant Workers: The Double Burden of the Two Viruses. *Liebertpub.Com*, *34*(6), 249–250. https://doi.org/10.1089/apc.2020.0092

Luo, W., Majumder, M., Liu, D., Poirier, C., Mandl, K., Lipsitch, M., & Santillana, M. (2020). *The role of absolute humidity on transmission rates of the COVID-19 outbreak*. https://doi.org/10.1101/2020.02.12.20022467

Ma, Y., Zhao, Y., Liu, J., He, X., Wang, B., Fu, S., Yan, J., Niu, J., Zhou, J., & Luo, B. (2020). Effects of temperature variation and humidity on the death of COVID-19 in Wuhan, China. *Science of the Total Environment*, *724*, 138226. https://doi.org/10.1016/j.scitotenv.2020.138226

Manoj, M. G., Satheesh Kumar, M. K., Valsaraj, K. T., Sivan, C., & Vijayan, S. K. (2020). Potential link between compromised air quality and transmission of the novel corona virus (SARS-CoV-2) in affected areas. *Environmental Research*, *190*, 110001. https://doi.org/10.1016/j.envres.2020.110001

Martin, C. A., Jenkins, D. R., Minhas, J. S., Gray, L. J., Tang, J., Williams, C., Sze, S., Pan, D., Jones, W., Verma, R., Knapp, S., Major, R., Davies, M., Brunskill, N., Wiselka, M., Brightling, C., Khunti, K., Haldar, P., & Pareek, M. (2020). Socio-demographic heterogeneity in the prevalence of COVID-19 during lockdown is associated with ethnicity and household size: Results from an observational cohort study. *EClinicalMedicine*, *25*, 100466. https://doi.org/10.1016/j.eclinm.2020.100466

Menard, S. (2002). *Applied logistic regression analysis*. Sage publications.

Merler, S., & Ajelli, M. (2010). The role of population heterogeneity and human mobility in the spread of pandemic influenza. *Proceedings of the Royal Society B: Biological Sciences*, *277*(1681), 557–565. https://doi.org/10.1098/rspb.2009.1605

Mittal, R., Meneveau, C., & Wu, W. (2020). A mathematical framework for estimating risk of airborne transmission of COVID-19 with application to face mask use and social distancing. *Physics of Fluids*, *32*(10). https://doi.org/10.1063/5.0025476

Mogi, R., & Spijker, J. (2020). *The influence of social and economic ties to the spread of COVID-19 in Europe*. https://osf.io/preprints/socarxiv/sb8xn/

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. *SIAM Journal on Computing*, *24*(2), 227–234. https://doi.org/10.1137/S0097539792240406

Neter, J., Wasserman, W., & Kutner, M. (1989). *Applied linear regression models*. http://www.sidalc.net/cgi-bin/wxis.exe/?IsisScript=LIBRO.xis&method=post&formato=2&cantidad=1&expresion=mfn=019124

Peterson, R. A., & Cavanaugh, J. E. (2020). Ordered quantile normalization: a semiparametric transformation built for the cross-validation era. *Journal of Applied Statistics*, *47*(13–15), 2312–2327. https://doi.org/10.1080/02664763.2019.1630372

Pullano, G., Valdano, E., Scarpa, N., Rubrichi, S., & Colizza, V. (2020). Evaluating the effect of demographic factors, socioeconomic factors, and risk aversion on mobility during the COVID-19 epidemic in France under lockdown: a population-based study. *The Lancet Digital Health*, *2*(12), e638–e649. https://doi.org/10.1016/S2589-7500(20)30243-0

Rowan, T. H. (1990). Functional Stability Analysis of Numerical Algorithms, 1990. *University of Texas at Austin*.

Rundle, C., Presley, C., Militello, M., … C. B.-J. of the A., & 2020, U. (2020). Hand hygiene during COVID-19: recommendations from the American contact dermatitis society. *Elsevier*. https://www.sciencedirect.com/science/article/pii/S0190962220322568

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*. https://projecteuclid.org/euclid.aos/1176344136

Steen, R., Hontelez, J., … O. M.-A. (London, & 2019, U. (2020). Economy, migrant labour and sex work: interplay of HIV epidemic drivers in Zimbabwe over three decades. *Ncbi.Nlm.Nih.Gov*. https://www.ncbi.nlm.nih.gov/pmc/articles/pmc6415983/

The New York Times. (2020). *GitHub - nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the US* https://github.com/nytimes/covid-19-data

Tibshirani, R. (1996). Regression Shrinkage and Selection Via the Lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Walter, L., Medicine, A. M.-W. J. of E., & 2020, U. (2020). Sex-and Gender-specific Observations and Implications for COVID-19. *Ncbi.Nlm.Nih.Gov*. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7234726/

Wang, J., Tang, K., Feng, K., & Lv, W. (2020). High Temperature and High Humidity Reduce the Transmission of COVID-19. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3551767

Wu, A. G., & Lipner, S. R. (2020). A potential hidden reservoir: The role of nail hygiene in preventing transmission of COVID-19. In *Journal of the American Academy of Dermatology* (Vol. 83, Issue 3, pp. e245–e246). https://doi.org/10.1016/j.jaad.2020.05.119

Xiong, C., Hu, S., Yang, M., Luo, W., & Zhang, L. (2020). Mobile device data reveal the dynamics in a positive relationship between human mobility and COVID-19 infections. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(44), 27087–27089. https://doi.org/10.1073/pnas.2010836117

Yeo, I. N. K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4), 954–959. https://doi.org/10.1093/biomet/87.4.954

Zambom, A. Z., & Kim, J. (2018). Consistent significance controlled variable selection in high-dimensional regression. *Stat*, *7*(1), e210. https://doi.org/10.1002/sta4.210

Zhang, G. Q., Pan, H. Q., Hu, X. X., He, S. J., Chen, Y. F., Wei, C. J., Ni, L., Zhang, L. P., Cheng, Z. S., & Yang, J. (2020). The role of isolation rooms, facemasks and intensified hand hygiene in the prevention of nosocomial COVID-19 transmission in a pulmonary clinical setting. *Infectious Diseases of Poverty*, *9*(1), 104. https://doi.org/10.1186/s40249-020-00725-z

Zhang, Z., Xue, T., & Jin, X. (2020). Effects of meteorological conditions and air pollution on COVID-19 transmission: Evidence from 219 Chinese cities. *Science of the Total Environment*, *741*, 140244. https://doi.org/10.1016/j.scitotenv.2020.140244

Appendix: Correlation table for period from 4 to 17

	TotalPop	RatioForeign	CitizenRat	NonCitizenRat	M_F_Rat	X18_25Rat	Pacific	Income	IncomePerCap	Poverty	ChildPoverty	Professional	Service	Office	Construction	Production	Drive	Carpool	Transit	Walk	OtherTransp	WorkAtHome	MeanCommute	EmployedPerc	PrivateWork	PublicWork	SelfEmployed	FamilyWork	diversity	target
TotalPop	1.00																													
RatioForeign	0.22	1.00																												
CitizenRat	-0.23	-0.31	1.00																											
NonCitizenRat	0.23	0.30	-1.00	1.00																										
M_F_Rat	-0.14	0.06	-0.16	0.16	1.00																									
X18_25Rat	0.10	0.04	-0.13	0.13	0.06	1.00																								
Pacific	0.17	0.11	-0.27	0.27	0.31	0.06	1.00																							
Income	0.35	0.13	-0.20	0.20	0.13	-0.21	0.04	1.00																						
IncomePerCap	0.43	0.11	-0.01	0.01	-0.09	-0.24	-0.04	0.90	1.00																					
Poverty	-0.49	-0.14	-0.04	0.04	-0.15	0.25	-0.02	-0.72	-0.67	1.00																				
ChildPoverty	-0.48	-0.15	-0.04	0.04	-0.20	0.12	-0.07	-0.71	-0.65	0.95	1.00																			
Professional	0.01	-0.05	-0.02	0.03	-0.16	-0.09	-0.05	0.60	0.66	-0.03	-0.08	1.00																		
Service	-0.36	-0.10	0.11	-0.12	-0.01	0.10	0.02	-0.45	-0.43	0.73	0.70	0.03	1.00																	
Office	-0.12	-0.21	0.01	0.00	-0.22	-0.18	-0.02	0.00	-0.01	0.24	0.26	0.30	0.38	1.00																
Construction	-0.63	-0.16	-0.13	0.13	0.27	-0.14	0.07	-0.35	-0.51	0.55	0.55	-0.14	0.53	0.29	1.00															
Production	-0.66	-0.22	-0.01	0.02	0.05	-0.04	-0.13	-0.50	-0.57	0.63	0.66	-0.16	0.36	0.20	0.65	1.00														
Drive	-0.53	-0.38	0.19	-0.19	-0.13	-0.24	-0.22	-0.22	-0.25	0.36	0.40	0.14	0.23	0.45	0.48	0.63	1.00													
Carpool	-0.46	-0.01	-0.30	0.30	0.17	0.02	0.20	-0.29	-0.42	0.59	0.58	0.03	0.52	0.26	0.70	0.56	0.15	1.00												
Transit	0.19	0.06	-0.19	0.20	-0.13	0.08	0.00	0.42	0.47	0.12	0.09	0.68	0.28	0.24	-0.07	-0.08	-0.16	0.13	1.00											
Walk	-0.25	-0.03	0.08	-0.08	0.05	0.29	0.00	0.02	0.09	0.43	0.36	0.49	0.53	0.13	0.16	0.20	-0.07	0.31	0.64	1.00										
OtherTransp	-0.15	0.03	-0.15	0.14	0.03	0.13	0.17	-0.02	0.03	0.48	0.42	0.43	0.53	0.30	0.26	0.16	-0.02	0.40	0.54	0.64	1.00									
WorkAtHome	-0.09	-0.02	-0.07	0.07	0.00	-0.19	0.11	0.39	0.39	0.08	0.04	0.69	0.26	0.45	0.15	-0.06	0.01	0.22	0.51	0.43	0.50	1.00								
MeanCommute	-0.10	-0.04	-0.12	0.12	-0.09	-0.33	-0.07	0.41	0.31	0.07	0.10	0.54	0.21	0.42	0.28	0.13	0.23	0.25	0.55	0.21	0.29	0.54	1.00							
EmployedPerc	-0.01	-0.02	-0.29	0.30	-0.01	-0.05	-0.08	0.64	0.60	-0.22	-0.24	0.69	-0.21	0.24	-0.06	0.05	0.16	0.04	0.53	0.29	0.20	0.44	0.37	1.00						
PrivateWork	-0.12	-0.17	-0.07	0.07	-0.11	-0.28	-0.21	0.12	0.15	0.15	0.19	0.36	0.14	0.45	0.20	0.47	0.46	0.20	0.35	0.22	0.19	0.35	0.38	0.52	1.00					
PublicWork	-0.51	-0.14	0.04	-0.04	0.04	0.21	0.07	-0.16	-0.25	0.58	0.50	0.34	0.56	0.24	0.44	0.26	0.26	0.46	0.26	0.49	0.47	0.27	0.28	0.00	-0.24	1.00				
SelfEmployed	-0.37	0.03	-0.04	0.04	0.07	-0.20	0.15	-0.04	-0.04	0.37	0.36	0.28	0.43	0.34	0.54	0.28	0.24	0.49	0.14	0.34	0.43	0.56	0.36	0.10	0.15	0.33	1.00			
FamilyWork	-0.30	-0.07	0.04	-0.05	0.00	-0.07	0.04	-0.09	-0.11	0.33	0.31	0.19	0.40	0.33	0.39	0.28	0.26	0.35	0.13	0.31	0.30	0.33	0.25	0.05	0.18	0.32	0.45	1.00		
diversity	0.62	0.30	-0.34	0.34	-0.04	0.23	0.22	0.12	0.14	-0.27	-0.27	-0.20	-0.28	-0.39	-0.46	-0.52	-0.63	-0.19	-0.01	-0.26	-0.12	-0.29	-0.23	-0.24	-0.42	-0.23	-0.44	-0.31	1.00	
target	0.02	-0.04	-0.03	0.03	0.01	0.10	0.01	-0.10	-0.12	0.09	0.09	-0.03	0.04	-0.01	0.03	0.13	0.08	0.09	-0.02	0.03	-0.06	-0.09	-0.08	0.04	0.04	0.04	-0.04	0.02	0.02	1