Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Modeling mobility, risk, and pandemic severity during the first year of COVID

Alexander Gilgur, Jose Emmanuel Ramirez-Marquez
Stevens Institute of Technology, Hoboken, NJ 07030, United States of America

ARTICLE INFO

Keywords:
Mobility
Pandemic
Vulnerability
Coronavirus
COVID-19
Severity metrics
Socioeconomic data
Data For Good
OxCGRT

ABSTRACT

During the COVID-19 pandemic, most US states have taken measures of varying strength, enforcing social and physical distancing in the interest of public safety. These measures have enabled counties and states, with varying success, to slow down the propagation and mortality of the disease by matching the propagation rate to the capacity of medical facilities. However, each state's government was making its decisions based on limited information and without the benefit of being able to look retrospectively at the problem at large and to analyze the commonalities and the differences among the states and the counties across the country. We developed models connecting people's mobility, socioeconomic, and demographic factors with severity of the COVID pandemic in the US at the County level. These models can be used to inform policymakers and other stakeholders on measures to be taken during a pandemic. They also enable in-depth analysis of factors affecting the relationship between mobility and the severity of the disease. With the exception of one model, that of COVID recovery time, the resulting models accurately predict the vulnerability and severity metrics and rank the explanatory variables in the order of statistical importance. We also analyze and explain why recovery time did not allow for a good model.

1. Introduction

It has been anecdotally and scientifically established since the early days of the COVID-19 pandemic that social and physical distancing, as well as personal protection equipment (PPE) and general hygienic practices, all contributed to people's safety. Yet for people working in stores, restaurants, at gas stations; people working in the taxi, ride-sharing, and delivery services; law enforcement officers; medics; itinerant workers, etc., — staying put is not an option. They have to travel during the workday, sometimes to other cities and counties away from their primary residence locations. Their exposure to coronavirus is presumably greater than that of people in the more affluent communities, who can afford to stay at home, ordering food and supplies to be delivered to them. Consequently, it stands to reason that in the poorer communities people would be more vulnerable to coronavirus than in the wealthier communities. This has been corroborated by multiple studies by CDC, WHO, and academic researchers (e.g., [1], or [2]). There have also been articles published describing use of mathematical methods in correlation analysis (e.g., [3]) and prediction (e.g., [4]) of spread of COVID-19.

Our goal is to formulate comprehensive predictive models to predict the severity of the disease and the risk to the individuals and the communities, based on which policymakers can develop holistic approaches to dealing with pandemics, targeting such approaches at reducing the severity of disease and the risks of contracting it in their constituent communities. To achieve this goal, we deemed it necessary not only to rely on the existing metrics such as disease propagation rate and mortality, but also to formulate new pandemic severity measures and to take into account the joint effects of mobility, socioeconomic, and demographic metrics at US County level and to be able to rank such metrics used as features in our models in the order of their relative importance for each of the pandemic severity measures.

In the early days of the pandemic, Oxford University came up with a way to quantify the strength of government response measures, creating a set of metrics for countries and for US States called OxCGRT (Oxford Covid Government Response Tracker). We explored the possibility of using OxCGRT composite indices to better understand how government actions influenced the propagation and severity of the disease. At the same time, the understanding that government-imposed restrictions may not be effective if the population does not observe them has led
us to formulating a new statistically and theoretically sound composite metric of mobility restraint – Mobility Restraint Index (MRI) – based on observed mobility data.

To measure COVID severity, we used the standard metrics — disease rate and mortality. However, these metrics do not point directly to risk and severity of the disease. To rectify this situation, we formulated three new statistically and theoretically sound metrics of disease severity and a probabilistic measure of risk — Recovery Time, Disease Severity Metric, and Risk. Details are presented in Section 3 and in Appendices.

Our models are built based on data from Data For Good Humanitarian Data Exchange (HDX), US Census Bureau, Bureau of Economic Analysis, Johns Hopkins Coronavirus Resource Center, and Oxford Covid Government Response Tracker (OxCGRT) to understand the relative importance of the metrics describing people’s mobility, socioeconomic status, and governmental responses in the United States for the risk of contracting COVID-19. Our intent was to make a step towards furthering our understanding of community resilience in a disaster at the scale of a deadly pandemic.

The rest of this paper is organized as follows. Section 2 is literature review. In Section 3, we describe our hypothesis and the methodological details, including the data sources, the modeling approach, and the exploratory data analysis. Section 4 describes the modeling results with regard to model evaluation and feature importance, while Section 5 is where we discuss the results and outline directions for further research opened by this work. In Section 6, we draw the final conclusions.

2. Literature review

2.1. Epidemiology

Epidemiology is defined in [5] as “the study of occurrence and distribution of health-related events, states and processes in specified populations, including the study of the determinants influencing such processes, and the application of such knowledge to control relevant health problems”. To understand how disease severity can be controlled, we need to understand the models used by epidemiologists. Two primary models are currently in use — Susceptible–Infected–Susceptible (SIS) and Susceptible–Infected–Recovered (SIR). Ridenhour et al. [6] provide details of these models, while [7] describes it at a more general level. Leung et al. [4] demonstrate how epidemiological models can be used to predict near-term transmissibility of COVID-19. Their paper is singularly important in that Hong Kong digital transaction metadata have been made available to the researchers. Due to privacy concerns, we have not been able to avail ourselves of such data in the United States. However, there are several authoritative sources of data on COVID propagation in the United States, most notably Johns Hopkins University [8], which adhere to the US privacy requirements, and which we used in our study. Ritchie et al. [9] underlines the importance of keeping in mind that the most ubiquitous derived metrics of disease severity, such as case rate and mortality, are relative; therefore the denominator (population or number of infected people, respectively) is very important.

2.2. Socioeconomic conditions

Socioeconomic conditions, such as income level, will have an effect on the individual's mobility: we assume that mobility will be different in the wealthy and upper-middle-class communities vs. in the lower-income and poor communities. There is a vast amount of research on measures of socioeconomic conditions, including a variety of composite metrics. We are interested in a link between socioeconomic inequality and objective well-being conditions between communities. Schenkmam and Bousquat [10] draw parallels between socioeconomic inequity and health outcomes, measured as infant mortality and life expectancy at birth. This analysis was conducted at the country aggregation level, and connected “health production (access, coverage and prevention) and intersectoral variables: demographic, socioeconomic, governance and health risks”. This work accounts for the cultural traits and physical and financial resources. The authors are focusing on health-related financial resources and not looking at the more generalizable socioeconomic metrics like personal income and its distribution, which are in the focus of our analysis: we are interested in the general wealth of the community and its effect on mobility and COVID vulnerability. An important finding is reported by Jung et al. [11]: “…there were disparities in excess mortality depending on community socioeconomic positions”. The authors also point out the importance of keeping track of the second-tier effect of COVID:

Patients with severe illness who visit the ED (emergency department) may also experience delays in proper treatment, which are due to a lack of resources resulting from the consumption of many medical resources, such as beds, intensive care units (ICUs) and medical staff, in the course of responding to the COVID-19 outbreak.

With regard to personal income and its distribution, we found two important papers: [12,13]. The former performs a thorough analysis of personal income distribution (PID) and concludes that, while the income bins used in computing Gini index have been historically too wide to enable an accurate estimation, “relevant overall PIDs (personal income distribution) evidence that the fixed income hierarchy has been observed from the very beginning of the CPS (US Census Current Population Survey)”. The latter proposes that power position of the individual, measured as the number of their subordinates, is the strongest determinant of the individual’s personal income. Extending the logic of comparing a city with a person proposed by de Oliveira Capela and Ramirez-Marquez [14], it can be hypothesized that the influence of a county or a city can be expected to have an effect on its average personal income.

2.3. Mobility

Multiple researchers report a relatively strong connection between mobility and propagation of the pandemic. Thus, [15] used a parametric model (multiple regression) to investigate the effect of mobility habits on the spread of COVID-19 in Italy. They show $R^2 \approx 0.42$ for their best linear model. While the authors correctly say that this is not an unusual value for $R^2$, this means that their model only explains 42% of the variance in the COVID spread metric that they aimed to predict. The authors use a “mobility habits” index for each region in Italy, which they calculate as the product of region’s population, average daily mobility rate prior to COVID-19, and percentage of variation due to the pandemic. It is similar to the mobility metrics that we use in our research: change in the range of daily movement. The average daily mobility rate (ADMR) is estimated as “percentage of residents in the … region who during a day make at least one trip, for whatever purpose, with the exception of pedestrian trips shorter than 5 min.…” This is similar to the percentage of limited-mobility residents that we use as one of the two mobility metrics we use in our work.

On the other hand, Banerjee and Lian [16] propose a model for predicting the spread of COVID based on mobility and mask mandate information. Their work is important for our research: by using nonparametric model (LSTM neural network), they have achieved predictions significantly better than obtained with the standard ARIMA method. However, the exceedingly low values of average percent difference that the authors achieved for New York (0.5%) correspond to $R^2 = 0.995$, which may point to overfitting. On the other hand, their results for Florida were generally acceptable, with average percent difference of $\approx 5\%$, or $R^2 = 0.95$. This tells us that nonparametric models of COVID spread by mobility and other factors have a high potential of explaining the variance in the predicted metric.
In [17], the author employs fixed effects models with correction for time trend. The author goes on to identify the modes of transportation and types of assembly that have a positive or a negative effect on \( R_c \), the reproduction rate of the disease. The author points out that “(t)he mobility measures provided are all highly correlated, making individual inferences on each impossible if all are included within the same model”. Therefore, the author goes on to estimate the independent effect of each of the mobility measures (retail, grocery, parks, transit, workplace, and residence) on \( R_c \) using a log-linear model and concludes that a significant reduction of mobility outside residence is needed to achieve \( R_c = 0.7 \). The author reports adjusted \( R^2 \) between 0.38 and 0.70. In this case, given the large sample sizes and the relatively small numbers of parameters, the difference between Adjusted \( R^2 \) and \( R^2 \) can be deemed insignificant. An important observation made by the author about the need to account for the incubation period of the disease is that a significant reduction of mobility outside residence is needed to prevent the disease from spreading.

Leung et al. [4] develop a nowcasting framework predicting COVID reproduction rate accounting for mobility, mixing of people, their ages, and other parameters. This work made it possible “to accurately track the local effective reproduction number of COVID-19 in near real-time”, eliminating the 9-day delay between infection and reporting of cases. The authors do not provide an estimation of their model fit, however, other than graphical representations of observed data and their predictions with confidence intervals. The fit is not great, but may be considered acceptable. In the spirit of [4,17,18] demonstrates the possibility to forecast COVID-related changes in mobility in six different categories (Retail and Recreation, Grocery and Pharmacy, Parks, Transit Stations, Workplaces, and Residential). The author demonstrates that, with the exception of Residential, all categories experienced a reduction in mobility in the Philippines during the early months of 2020, when the country experienced an onset of COVID-19.

2.4. Government response to pandemics

Government response to the pandemics describes state government policies and is very important for modeling disease severity, propagation rate, and risk to the individuals and communities. Pak et al. [19] conduct a regression analysis of public trust and compliance with government measures to stem the spread of COVID-19. They conclude that (i) stricter government restrictions lead to an increase in compliance and (ii) “(a)mong political regimes, higher levels of public trust significantly increase the predicted compliance as stringency level rises in authoritarian and democratic countries”. Díaz-Castro et al. [20]’s analysis is different from [19] in that the authors investigate the differences in compliance among states of one country (Mexico), which represents a similar approach to ours for the United States. The authors brought together the University of Oxford Coronavirus Government Response Tracker (OxCGRT), population size and density, poverty, and mobility, and calculated the doubling time of COVID-19 cases as the response variable of their model. The parallel with our work is obvious. The authors use a linear model to predict doubling time, putting together the listed metrics. The model suggests that “income per capita emerged as a strong significant positive predictor of the doubling time of COVID-19 cases”.

At the same time, it appeared that mobility and prevalence of poverty, too, are positively correlated with the doubling time — reducing COVID spread rate. On the surface, this seems paradoxical. The authors explain this by going deeper into the socioeconomic and demographic situation in the states with higher poverty rate. We believe that in this case, linear model with strongly correlated features (GDP and Poverty; Population and its Density; an unknown relationship between Poverty and Mobility, etc.) should not have been used directly due to the multicollinearity, as it may lead to paradoxical effects (e.g., Simpson’s paradox, [21,22]): collinear features interact in ways that appear strongly counter-intuitive. That said, main effects are not susceptible to Simpson’s paradox. Therefore, the authors’ statement that “increased mobility was significantly associated with more rapid viral transmission” is statistically sound.

On the other hand, the use of doubling time as the response variable in a linear regression should be taken very cautiously: time between independent events does not follow the Gaussian distribution, resembling either an Exponential, a Gamma, or a Power-Law distribution. Due to these reasons, unless confidence intervals of the prediction are not important, linear regression should not be used in this kind of analysis. Overall, the work by the authors of [20] is highly innovative and inspirational. The authors of [23] explain that the spread of a pandemic should be modeled using the S–I–R (Susceptible–Infected–Recovering) model, not by a linear formula. This, too, may be responsible for the logical discrepancies reported by Díaz-Castro et al. [20]. This is corroborated by Buckman et al. [24], who extend the S–I–R model to account for the deviations from steady-state conditions and compute the \( R_0 \) (basic infection rate) and \( R_e \) (effective infection rate) to predict the peaks in the numbers of COVID cases in the US, Brazil, China, and Italy.

2.5. Community resilience and cohesion during COVID-19

It takes the entire community to stop a community-wide disaster. Coordination and willingness to undertake the necessary measures is key to stopping it. Therefore cohesion and resilience are very important factors in this research.

Henry and Ramirez-Marquez [25] developed a general framework of measuring system resilience, where the system’s performance function, \( F(t) \), shifts level from a stable value due to a disturbance and after a time recovers to a new stable value. The difference between the original stable value and the disturbed level, divided by the difference between the recovered level and the disturbed level is proposed as a measure of community resilience: \( M_R = \frac{F(t) - F(t_d)}{F(t) - F(t_r)} \), where \( t_d \) is the last measurement of \( F(t) \) before the disturbance; \( t_r \) is the time of stabilization at the disturbed state; \( t_c \) is the time of stabilization at the recovered level of \( F(t) \). Gilgur and Ramirez-Marquez [26] further developed the measurement of community resilience. The authors take the idea of resilience as a process and view its metric as a property of the community, proposing to measure it by the sensitivity of stabilization time to the size of the disturbance. They then propose a self-correcting Bayesian method to measure community resilience as the inverse of the sensitivity of the recovery time to the severity of the disturbance. This approach is demonstrated to be viable in a set of longitudinal simulations, where one community is assumed to be taking multiple disturbances. We believe that, when multiple communities are hit with the same disturbing event, such as COVID-19 pandemic, the same approach can be applied as well.

Jewett et al. [27] is one of the most recent publications on the topic of community resilience during the COVID-19 onset. They outline the current state of research into social cohesion and resilience during the pandemic. The authors describe equity, sustainable development, and separation between the public and the government as factors affecting social cohesion. However, no data are proposed in support of the analysis in this paper. It is intuitive that communication, especially risk communication, can be instrumental in building community resilience. In [28], the authors examine the importance of public awareness of risk and the role of public participation in risk communication and resilience building.

The authors warn:

A community’s resilience to tragedies strongly depends on the communication surrounding them before, during, and after crises. ... Understanding of this process, however, is limited, and efforts to improve public awareness of hazards and removing people from harm’s way have proven ineffective.
They further explain that “...social media should be viewed as complementing traditional media, not as operating in opposition to traditional media (Jin and Liu 2010, op.cit., p. 199)”. We would like to add to this that even when social media operates with traditional media, they should be viewed as independent entities. This is of importance for our work, as it pertains to the analysis, among other features, of the effect of state government measures in response to COVID on its contagion and mortality.

2.6. Model quality baselining

Our research involved creating and modeling new metrics. It is important to see what portion of the variance in these metrics our models explain, and whether they are on par with models reported in the literature by other researchers. This question is answered by baselining the $R^2$ of the models. Based on the literature review and our models, we have compiled Table 1. Importantly, the results presented in it are only to be used to qualitatively establish that our models are at least as good as the models we found in the literature. As we see from Table 1, [16]’s LSTM neural network model quality is the best of the compared models, albeit, as we discussed in Literature Review, it may be an overfit for some communities (New York, 0.995). With our approach, the model quality exceeds two out of three cited models for 4 out of 5 metrics, with one metric in the same ballpark as [17] and better than [15].

3. Methodology

In this section, we describe how we connected the predicted metrics – traditional (COVID rate, Mortality) and derived (Recovery Time, Severity, and Risk) – with government measures and mobility, as well as demographic and socioeconomic factors. Derivation of all metrics is outlined in Fig. 1 and detailed in the Appendices.

3.1. Hypothesis and working model

Our hypothesis is that people’s socioeconomic status, population density, government measures to contain the pandemic, as well as mobility have an impact on the risk of contracting the virus, as well as severity of the illness. We validate it by modeling and analyzing the quality of our models. Five categories of metrics interact in complex feedback and feed-forward relationships in our models: (i) Government Response; (ii) Socioeconomic Metrics; (iii) Population Mobility (iv) Population Density Metrics; (v) COVID Metrics. We put together data from several sources, quantifying people’s mobility, socioeconomic conditions, population dynamics, and state government responses to COVID-19. (See Fig. 2 for illustration.) We predict COVID Rate, Mortality, Disease Severity Metric, Risk, and Recovery Time as functions of mobility, government measures, socioeconomic, and demographic metrics. Details on the definitions of dependent and independent metrics are presented in Section 3.5. Derivations are provided in the Appendix.

3.2. Data sources

In this research, we used the following publicly available data sources:

1. The Humanitarian Data Exchange (HDX) is an online assembly of data, contributed by industry, academia, and governments, enabling researchers and policymakers to analyze a number of important metrics, such as, during the COVID-19 pandemic, movement range data.
   URL: https://bit.ly/3qmYvwb

2. Johns Hopkins Coronavirus Data. The Johns Hopkins Coronavirus Resource Center (CRC) is a continuously updated source of COVID-19 data and expert guidance. CRC team “…collect and analyze the best data available on cases, deaths, tests, hospitalizations, and vaccines to help the public, policymakers, and healthcare professionals worldwide respond to the pandemic”.
   URL: https://coronavirus.jhu.edu/data;
   Github: https://bit.ly/3qnN1Z0

3. Personal Income Data by US County. This dataset was downloaded from Bureau of Economic Analysis.
   URL: https://bit.ly/2OtgbJf

4. Population and Land Area of US Counties. These data were downloaded from the US Census Bureau site
   URL: https://bit.ly/2PvVdK4

5. Oxford Covid Government Response Tracker (OxCGRT) was created by Blavatnik School of Government at the country level during the early days of the pandemic. Later a US-specific tracker was developed. We downloaded it and used in our research.
   URL: https://bit.ly/38foBuu;
   Github: https://bit.ly/38foBuu

![Fig. 1. Origins of the New metrics.](image-url)
variables are influenced by the selected features, it enables researchers to evaluate the relative importance of the features for each of the dependent variables. To build an independent model for each of the dependent variables, we need to verify that they are uncorrelated with each other. This is done in Section 3.5.

3.5. Metrics

In this work, we analyzed the possibility of modeling COVID-19 metrics based on government measures, as well as mobility, socioeconomic, and demographic factors. In this subsection, we list the metrics available to us, derived metrics, and how we used each of these metrics in our models.

Derivations, mathematical expressions, and other details pertaining to the metrics are provided in the Appendices.

3.5.1. Available and standard derived metrics

Johns Hopkins University Coronavirus Research Center [8] provides data on the total daily number of confirmed COVID cases (cumulative and daily) and total number of deaths in the US at the county granularity level. We call COVID rate and COVID mortality standard derived metrics because they are computed and used in this research in the same way as, e.g., defined in [9].

We used two standard derived metrics in our work:

1. COVID RATE. COVID rate is an important measure of disease propagation in a county. It expresses the proportion of cases in the population. It was computed as

\[ \kappa = \frac{N_c}{N_p}, \]  

where \( N_c \) is the number of cases in the county, and \( N_p \) is the county’s population count.

2. COVID MORTALITY. COVID mortality is an aggregate measure of severity of the disease and the state of preparedness, or lack thereof, of the county’s medical facilities, supplies, and personnel, including first responders and hospital workers, for handling the pandemic. It was computed as the number of deaths divided by the number of cases: \( \eta = \frac{N_d}{N_c} \), where \( N_d \) is the day’s death toll.

3.5.2. New (GRM) derived metrics

To further our understanding of the diverse aspects of the pandemic represented by Johns Hopkins University’s Coronavirus Research Center (CRC)’s data, and the impact of mobility and socioeconomic metrics on the severity of the disease and people’s risk, we developed the following new metrics of COVID severity and risk. The following metrics were derived in this work:

Recovery time (average duration of being ill).

Disease recovery time is a very important metric for community planning and preparations for pandemics and other health-related disasters. It is an important indicator of disease severity and a critical
element in estimating capacity and personnel requirements for medical treatment facilities and the impact of the disease on the economy. We used Eq. (2) for the average time to recover (duration of being ill)

\[ T = \frac{N_c - N_d}{\frac{d(N_c - N_d)}{dt}}. \]  

(2)

where \( N_c \) is the number of cases; \( T \) is the average recovery time (duration of the person’s being ill), \( N_d \) is the number of deaths. Then \( d(N_c - N_d)/dt \) represents recovery speed, and \( N_c - N_d \) is the concurrent number of recovered COVID patients. For derivation, please refer to Appendix B.3.1. Hereinafter Recovery Time, Duration of Being Ill, as well as days_sick will be used interchangeably.

**Disease Severity Metric.**

To streamline the analysis and modeling, and to avoid confounding introduced by correlated features, we formulated a Disease Severity Metric – a linear combination of these three measures – total number of cases, new cases (arrival rate), and death toll. We define it as

\[ DSM(t) = \sqrt{\sum_{i \in PC_s} P_i(t)^2}, \]

(3)

where \( DSM(t) \) is the value of Disease Severity Metric at timestamp \( t \), \( PC_s \) is the set of principal components (see below), and \( P_i(t) \) is the \( i \)th eigenvalue (size of the \( i \)th principal component) at timestamp \( t \). The rationale behind this approach is as follows. As we explained in Appendix B.1, Johns Hopkins University’s CRC provides three important standard derived metrics — total number of cases, new cases (arrival rate), and death toll, and they are pairwise correlated. To resolve the correlations, we used PCA. (See Appendix A.2 for details. For derivation, please refer to Appendix B.3.2. Disease Severity Metric is a statistically sound composite metric combining the signal from new COVID cases, total COVID cases, and number of deaths. It carries as much information as its component metrics, but it removes the redundancy intrinsic in multicollinearity. For these reasons, we are using the Disease Severity Metric (DSM), rather than its elements, as an output variable of our model in this study.

**Covid risk**

Risk is the metric bringing together the severity of the disease and the probability of getting the disease.

\[ R = D \ast \kappa, \]

(4)

where \( R \) is the risk; \( D \) is severity measured by the Disease Severity Metric. In turn, \( \kappa \) is COVID rate computed as defined by Eq. (1). For details, please refer to Appendix B.3.3.

**Mobility restraint index**

In Section 3.1, we stated our hypothesis that mobility, along with socioeconomic status, population density, and government measures, have an effect on the severity and propagation rate of the disease. We have two aggregate metrics of daily mobility during COVID, proportion of each county’s population that stayed within the same Level 16 Bing tiles (600 × 600 m squares) during the day and range of movement change measured as the relative change in the number of boundaries of Level 16 Bing tiles that each individual crossed during the day. They describe two different aspects of mobility restraint within each US county. To combine the information from these two metrics and eliminate redundancy, we formulate a metric of mobility restraint exercised by the population. Like Disease Severity Metric, Principal Component Analysis (PCA) has been used to formulate this metric, and like DSM, we defined MRI using equation similar to Eq. (3):

\[ MRI(t) = \sqrt{P_1(t)^2 + P_2(t)^2}, \]

(5)

where \( MRI(t) \) is the mobility restraint index (MRI) at time \( t \), \( P_1(t) \) is the first principal component at time \( t \), and \( P_2(t) \) is the second principal component at time \( t \), returned by PCA. Please see Appendix C.3 for details.

**Fig. 3. State Government Measures (OxCGRT data), Mobility, and Risk of COVID.**

**Table 2**

| StringencyIndex : Risk | -0.416 |
| StringencyIndex : Mobility Restraint Index | 0.143 |
| StringencyIndex : Disease Severity Metric | 0.081 |
| StringencyIndex : Risk | 0.273 |

**3.5.3. Exploratory data analysis. Important correlations**

We analyzed pairwise correlations of COVID metrics and population density, socioeconomic metrics (average personal income), state government response measures, and mobility. One of the important discoveries was that, while there are weak linear correlations (see Fig. 3), most relationships are characterized by one or two critical values of the driving variable: e.g., government-response Stringency Index has a critical value above which we observed a reduction in COVID rate and mortality (Fig. 4). On the other hand, income level has two critical values, falling at the Pew Research Institute’s boundaries of middle income, between which Mobility Restraint Index was at its highest value (Figs. 6 and 7).

**State government response measures and other metrics**

From Fig. 3, we see that there is a noticeable negative correlation between Stringency Index (hereinafter SI) and county COVID rate \((N_{cases}\_rate}/N_{residents})\), \( \rho = -0.42 \), and other relevant metrics. Because of the stronger correlations between Stringency Index (SI) and COVID-related, mobility, and socioeconomic metrics, we are focusing on it as the primary measure of government response. Table 2 provides the pairwise Pearson correlations between SI and some other metrics. The P values in Fig. 3 and in Table 2 are less than 0.001, meaning that the correlations are significant with confidence higher than 99.9%.

Note that correlations between SI and COVID Risk and DSM (i) are weak and (ii) should not be read as increases in government response stringency causing increase in disease severity or risk. It is reasonable to assume that in states where COVID was more advanced, governments took stricter measures against the disease.

The time shift described in Section 3.3 accounts for the delay in implementation of government measures and the normal time constant inherent in the system.
The images in the visualizations below show that a big part of the variance in the dependent variables (on the vertical axis of Figs. 4 and 5b) is not explainable by a single independent variable shown on the respective horizontal axis. However, it is clear that the relationships shown in these visualizations are nonlinear, and many have a critical point where the behavior of the dependent variable changes. We illustrated this by showing the vertical dotted lines at such points and the arrows to highlight the visible behavior at the lower and the higher values of the independent variables. From these illustrations, it is clear that linear modeling will not work well in describing and predicting the behavior of the metrics of interest, and we need to use nonparametric modeling (e.g., Random Forest Regression) to derive rigorous models. At the same time, the existence of critical points identified in these plots will help in informing the government policies and sensitivity analysis of the RFR models. We see from Fig. 4 that there is a nonlinear relationship between the Stringency Index and the rate of COVID spread \( \left( \frac{N_{\text{cases}}}{N_{\text{residents}}} \right) \), as well as Stringency Index and COVID mortality \( \left( \frac{N_{\text{deaths}}}{N_{\text{cases}}} \right) \).

Table 3

| Bucket   | \( Rate_{avg} \) | \( Rate_{p90} \) | \( Mortality_{avg} \) | \( Mortality_{p90} \) |
|----------|------------------|------------------|-----------------------|-----------------------|
| Below CV | 0.32             | 0.086            | 0.022                 | 0.047                 |
| At CV    | 0.031            | 0.079            | 0.031                 | 0.082                 |
| Above CV | 0.011            | 0.038            | 0.028                 | 0.083                 |

Note that there is a critical value of Stringency Index, at about 40% of the range, at which COVID-19 rate starts visibly decreasing. We shall call it rate-critical value of SI, \( SI_{rc} \). At the same time, COVID mortality \( \left( \frac{N_{\text{deaths}}}{N_{\text{cases}}} \right) \) kept increasing past \( SI_{rc} \), despite the increase in StringencyIndex (Fig. 4(b)). While a piecewise regression analysis would be needed to better identify the critical values and characterize the behavior of the metrics below, at, and above such values of the StringencyIndex, the focus of this paper is beyond bivariate correlation analysis. We are building a comprehensive predictive model of COVID risk and severity metrics, accounting for the other features and their interactions. That said, Table 3 provides the numerical estimates of the mean and \( p_{90} \) of rate and mortality below, at, and above the corresponding critical values (CVs) of the StringencyIndex. The numbers corroborate the visual observations.

The increase in COVID mortality at SI below its critical value is not to be interpreted as a statement that government actions were causing an increase in mortality — it is more reasonable to assume that restrictions imposed by governments were in response to the growing rate and mortality of the disease. However, this causality may go in both directions. Indeed, many actions taken by federal, state, and local governments in the US, especially in the early days of the pandemic, were made based on extremely limited information about COVID-19’s spread mechanisms and its contagiousness, under conditions of long test turnaround times and shortage of test kits. These decisions, and delays, often resulted in mistakes, leading to increased contagion and mortality. At a mortality-critical value of StringencyIndex, \( SI_{mc} \approx 60\% \), the tide began to turn.

**Mobility restraint index, COVID mortality, and risk**

From Eq. (4), it is clear that there is a negative nonlinear correlation between MRI and COVID risk. It is intuitively understandable: counties where people travel more frequently, and longer distances, are likely to have higher COVID risk than the lower-mobility counties, and we see it in Fig. 5(a). We also see from Fig. 5(b) that there is a clear critical value of MRI, at \( \approx 30\% \) of the range, above which COVID mortality starts going down.

**Mobility restraint index and personal income**

The relationship between personal income and mobility restraint due to COVID has turned out to be more complicated than we had expected. Thus, we see from Fig. 6 that middle-class counties had a higher Mobility Restraint Index than their wealthier and lower-income counterparts. Income data are strongly skewed, obscuring the low- and middle-income numbers for Mobility Restraint Index. To resolve it, we used the logarithmic transformation of the X axis in Fig. 6. It shows that MRI in the low-income counties increases with income, reaching its maximum in the middle-class range (Fig. 6 for reference), and going down at higher incomes. We see that the peak of Mobility Restraint Index (MRI) falls on the range which, according to the Pew Research Center \([30]\), corresponds to the upper middle class in the US (see Fig. 7). Table 4 shows the aggregate values of MRI in the three income buckets delineated by the vertical lines in Fig. 6.

We see that both the average and the high levels of MRI are higher in the middle-class counties than in the predominantly low- and high-income counties. Based on the analysis of Figs. 6 and 7, we conclude that there may be some sort of a causal relationship between mobility restraint and income, and middle-class counties have been more restrained in their mobility. A possible explanation is hypothesized as follows. Low income communities have low Mobility Restraint Index:
their members cannot afford to stay at home. Communities in the higher income range, approximately from the middle of the middle-class range ($35,000–$50,000), are characterized by the highest MRI values.

Circumstantially, we can connect this with (i) the closures and workforce reductions of many businesses considered non-essential during the COVID pandemic, leading to people not driving to and from work every day, and (ii) people being able to work from home during COVID.

Finally, for the upper middle class and the wealthy communities, Mobility Restraint Index goes down. Following the same logic, we can say that their mobility plays a role in their being in the high-income range. US Census data for personal income from the last pre-COVID year, 2019, were used in this analysis; the economic impact of the pandemic is outside the scope of this research.

3.6. Summary of COVID metrics and their factors

In Section 3.5, we made the following important findings: (i) There are many collinear metrics in our datasets, and it is important to be able to group them. Principal Component Analysis (PCA) provides one way of doing so. We used it to formulate Disease Severity Metric (DSM) and Mobility Restraint Index (MRI). (ii) There is a strong correlation between mobility restraint and risk of COVID defined as Eq. (4) — the product of COVID rate and Disease Severity Metric, which we defined in Section 3.5.2. (iii) There is also a strong non-linear correlation between socioeconomic conditions (average personal income) and mobility restraint. (iv) Nonlinear correlations exist between MRI and COVID rate, as well as MRI and COVID mortality. (v) Other variables, such as
personal income and population density interact with MRI, Stringency Index, and COVID metrics.

4. Results. Model evaluations and factor importance analysis

In this section, we build and evaluate descriptive models that can be used to predict COVID severity and propagation rate, to help inform state policies and identify local measures aimed at preventing the spread of pandemics. Such models can be used in a simulation environment to answer a variety of what-if questions. In their simplest form, they can be used to measure the relative importance of the factors that we have taken into consideration. We used the approach outlined in [29], with some modifications. Specifically, due to the new (GRM) features we derived (see Fig. 1), we did not need to perform any further dimensionality reduction: the set of features is relatively small. The Random Forest Regression (see Appendix A.1 for the rationale behind choosing this method) was fitted using the following features: (i) Mobility Restraint Index (MRI); (ii) Population Density; (iii) Personal Income; (iv) StringencyIndex; (v) ContainmentHealthIndex; (vi) EconomicSupportIndex.

We predicted the following COVID metrics: (i) COVID Rate; (ii) COVID Mortality; (iii) Average duration of being ill (Recovery Time); (iv) Disease Severity Metric (DSM); (v) COVID Risk

4.1. Model hyperparameters and evaluation procedure

In this analysis, we used the default RFR parameters from the sklearn.ensemble.RandomForestRegressor Python package. We increased the number of decision trees (evaluators) in the random forest to 50 (the RandomForestRegressor version we used had a default of 10. Beyond 50 we did not notice significant improvements in model performance). Loss function has been set to the L2 Norm — mean-square error (MSE), to maximize the models’ robustness to outliers and pivot points. We enabled bootstrapping to ensure that each decision tree in the random forest got a sufficient number of data points and did not limit the maximum number of features used in each tree. If we had obtained poor results in terms of model quality, we would have tuned these hyperparameters for each model.

To evaluate the models and to ascertain their ability to predict unseen data, we used a 3:1 train:test split of the data aggregated by county and day, making each data point correspond to a county (day) pair. As we discussed above, we used feature engineering techniques, such as time shifting, to avoid finding spurious correlations. With the understanding that relationships among the metrics would develop in time, and therefore leaving time outside the list of independent variables would add uncertainty to our models, we focused on finding the underlying fundamental relationships between our features and the predicted metrics. In the future, we are planning to add a temporal component to our research (see Section 5.2). Abstracting away the time enabled us to streamline the model quality evaluation by applying the standard 3:1 randomized split between the training and the testing datasets.

4.2. Model quality analysis

In order to estimate the degree of trust in the stack ranking of features performed by Random Forest Regression models, we must know how good the models are. Each of the COVID metrics we analyzed in this research uniquely quantifies the risk, severity, and other aspects of the pandemic. This makes it important to independently evaluate the models and identify the order of importance of its features, for each COVID metric. Using a 3:1 train:test split of the datasets for each model enabled us to ascertain that the models generalize sufficiently well. Table 5 summarizes our findings. As we can see from it, features that remained outside the scope of our work are responsible for 7% to 39% of the variance in the respective COVID metrics. Visual analyses of the best and the worst models as scatter plots of model prediction as a function of observed metric value tell us whether the model is good or we need to change the model hyperparameters, search for additional features, or make the conclusion that this direction is not going to be successful.

4.3. Model summary

To visualize model quality, we plotted model prediction as a function of observed values in a scatter plot. When the points fall more or less on the 45° line, it tells us that the bias in model prediction cannot be improved any more. E.g., in Fig. 8: (a) COVID Rate and (b) Recovery Time. We observe a good fit for COVID Rate, but we also see a clear bias for Recovery Time. In this case, the model for Recovery Time is underfitting the data: we do not have the missing features needed to explain the variance of this metric.

4.4. Feature analysis

A brief summary of feature importances is below.

COVID Rate: As we see in Table 5, the most important feature for COVID Rate is StringencyIndex, followed by ContainmentHealthIndex, Population Density, and Personal Income. MRI and EconomicSupportIndex are not as critical for COVID rate as the other four metrics.

Recovery Time: Of the metrics used in our model, MRI; StringencyIndex; and ContainmentHealthIndex are equally important, while Population Density, Personal Income, and EconomicSupportIndex play a significantly smaller role in this metric. Other factors, like COVID-readiness of the local hospitals, including availability of ICU beds,
COVID-specific diagnostic and treatment equipment and medication, as well as staffing and fatigue levels of the front-line medical personnel, supply of personal protective equipment (PPE), infected individuals’ age, medical history, immediate pre-COVID conditions, and other factors that are outside the scope of this study will be taken into consideration in followup research.

COVID Mortality: Here too, availability of rapid test kits, hospital personnel and ICU beds, supply of personal protective equipment (PPE), and other factors that are outside the scope of this study must be taken into consideration; while the model was not the worst fit at $R^2 = 0.76$, it still leaves 24% of the variance unexplained. Of the features used in modeling, Personal Income, ContainmentHealthIndex, and Population Density are the Top 3 predictors of COVID Mortality.

Disease Severity Metric: Notably, Population Density is the single most important feature of our model for Disease Severity Metric (DSM) — a linear combination of new cases, death toll, and total number of cases, followed with a large margin by MRI. Given that the model fit was good at $R^2 = 0.85$ — the model explained 85% of the variance in DSM, leaving only 15% to metrics that were out of scope of this research — it is reasonable to say that, despite well developed medical and emergency infrastructure in the high-density counties, the probability of contact with infected individuals has led to higher DSM values than in the rural (low-density) counties.

COVID Risk: Because of the way we defined risk in Eq. (4), it is no surprise that Population Density is the most important factor for this metric. What is surprising is that MRI has shifted down to the fifth position out of the six, displaced by ContainmentHealthIndex, StringencyIndex, and Personal Income. Intuitively, we believe two factors are at play here: (i) MRI does not account for how many county lines the individuals crossed during the day, and what the COVID situation was in the counties entered; and (ii) Personal Protective Equipment (PPE) was becoming available and mandated by the state and local regulations at about the same time as MRI was restabilizing (Fig. 18). These circumstances have confounded the effect of MRI on COVID Risk. Mathematically, considering that ContainmentHealthIndex and StringencyIndex are the most important features for COVID Rate, and Rate is one of the components of our definition of COVID Risk, it makes sense.

5. Discussion and directions of further research

We have identified the metrics related to propagation of the COVID-19 pandemic. We understood their interactions and, based on these interactions, developed composite metrics that work well to predict disease severity, rate, and mortality in a pandemic.

These metrics and disease characteristics have been put as features and target variables, respectively, through Random-forest Regression models to rank the features in order of their effect on each of the selected metric.

5.1. Observations and analysis

StringencyIndex, ContainmentHealthIndex, Personal Income, and Population Density, as well as Mobility Restraint Index are strong contributors to the COVID-related metrics. EconomicSupportIndex has negligible effect on any of the five COVID metrics we have considered in this study.

• Government measures – their StringencyIndex and Containment HealthIndex – have a strong effect on COVID rate — measured as $N_{cases}/N_{residents}$, along with (to a smaller degree) Population Density and Personal Income. Mobility Restraint Index (MRI) has negligible effect on COVID rate.

• Mortality — measured as $N_{deaths}/N_{cases}$ — is equally affected by Personal Income, ContainmentHealthIndex, and Population Density. Stringency Index (SI) and MRI, as well as EconomicSupportIndex each have $\approx 1/2$ of their effect. Of the metrics considered in this study, duration of being ill (days sick) needs the most work in the future. Mobility restraint has the biggest effect on it; however there are other factors that are outside the scope of this work.

• Disease Severity Metric (DSM) — a linear combination of new COVID cases, COVID death counts, and average duration of being ill – from diagnosis to second negative test — is primarily affected by Population Density and Mobility Restraint Index (MRI). Personal Income and Government measures have a smaller effect on this metric.

• COVID Risk – defined as the product of DSM and COVID rate – is primarily affected by Population Density, government health- and containment-related measures, and overall stringency of government measures. The effects of Personal Income and MRI are significantly smaller.

5.2. Further research

This work has posed important questions along the following axes:

• RECOVERY TIME As we saw in Section 4, COVID recovery time is affected by features that remained outside the scope of the current research.

• GOVERNMENT MEASURES AND MOBILITY We saw in Fig. 3 that the correlation between state governments’ StringencyIndex and MRI is very low at 0.14. Fig. 4 tells us that their relationship is very complex, with possible feedback causal loops, as well as other factors, including political and social events happening at the same time. We would like to better understand the factors at play, to enable state and local governments to form more effective policies and measures with the purpose of getting a more intuitive response from the population.

Fig. 8. Model prediction quality for COVID Rate (a) and for Recovery Time (b). In the ideal fit, all points fall around, and near, the 45° line. Local Deviations from this line usually point to bias, which often can be mitigated by expanding the feature search.
• PUBLIC AND SOCIAL MEDIA, GOVERNMENT SUPPORT, AND COMMUNITY RESILIENCE

Public media, and people’s communication through social media, increase people’s awareness of the risks of COVID. This serves as a feedback to people’s movement range, moderating mobility metrics. In turn, this reduces the risk of spread of coronavirus, but has the unintended consequence of risk to the livelihood and welfare of those whose income is not high enough to enable them to stay in place through the day (e.g., to work from home). To offset this, federal, state, and local governments have been taking economic support measures, ranging from anti-eviction regulations and rent control to stimulus packages and are designed to help the low-income individuals sustain their livelihoods and welfare during the pandemic. This works as a feedback loop. Its effect on community resilience is very important to understand in order to inform public policies and to foster resilient communities.

• TIME AS AN EXPLANATORY VARIABLE As we discussed in Section 4.1, time is an important parameter affecting how the relationships between the explanatory and the dependent variables evolve. Now that we have a way to reduce the scope of the modeling problems by highlighting the statistically important features via nonparametric modeling, we can incorporate time as an explanatory variable — either (i) by parametric modeling of the relationships we found in the current work and analyzing the time dependency of the parameters of such models, or (b) by incorporating modern developments in multivariate time series analysis and deep learning, as proposed, e.g., by Banerjee and Lian [16].

6. Conclusions

Fighting the deadly pandemic with science and data has been paying off. Consequently, many sources of authoritative publicly available data have been formed. To better understand the complexity of the processes happening during a pandemic, and to help formulate policies in the future, we have built predictive models explaining the joint impact of mobility, socioeconomic, and demographic variables on disease severity.

In this process, to measure disease severity, we used the standard metrics of disease rate and mortality, and also defined new statistically and theoretically justified metrics, Disease Severity Index (DSM), Average Recovery Time (Duration of Being Ill), and Risk. To streamline the analysis of the effects of mobility on COVID metrics, we formulated a metric that we called Mobility Restraint Index (MRI). This enabled us to (i) investigate the applicability of state-government response metrics formulated by Oxford University (OxCGRT) to predicting the outlook of the pandemic; (ii) determine the existence of critical values of OxCGRT composite metrics at which COVID spread slows down and patients’ chances of survival improve; (iii) build comprehensive models connecting socioeconomic, demographic, mobility, and government response metrics with pandemic-related measures, such as rate, mortality, severity, and risk; (iv) rank the factors in the order of their relative importance for each of the pandemic measure, identifying which features should be used in predicting each of the dependent variables. The new metrics and the models contribute to our understanding of the pandemic and open new opportunities for further research. We outlined some of these directions in Section 5.2.

CRediT authorship contribution statement

Alexander Gilgur: Equally contributed to the detailed development of the idea. Contributed the identification of data sources. Data engineering. Tool development. Implementation of the metrics and the algorithms. Analyses of the results. Jose Emmanuel Ramirez-Marquez: Contributed the idea, Advising, Editing. Reviewing during all stages of our work on this paper. Equally contributed to the detailed development of the idea.

Appendix A. Choice of modeling tools

A.1. Random Forest Regression

Random Forest Regression (RFR) is a predictive-modeling tool with the following advantages over other regression methods:

1. It is one of the standard machine-learning methods.
2. RFR does not make any assumptions about linearity of the model, distribution of the error terms, or absence of multicollinearity.
3. Feature importance is computed in RFR as part of the procedure.
4. RFR accounts for interactions between features.

Its weakness is its interpretability: while RFR computes feature importances, it does not estimate the sensitivity of the model to each of the features. This can be surpassed by following up with a sensitivity analysis or the more recent machine-learning interpretation tools (e.g., LIME or SHAP). For these reasons, we chose Random Forest Regression (RFR) for modeling. Other tree-based regression methods, e.g., Boosted Decision Trees, have the same advantages as RFR. However, in our experience, Random Forest methods tend to work better with high-volatility (noisy) data, whereas Boosted Decision Tree methods tend to be prone to overfitting noise, making models less likely to accurately predict new data.

A.2. Principal Component Analysis

For formulation of Mobility Restraint Index and Disease Severity Metric, we used Principal Component Analysis (PCA). It is a standard method for finding linear combinations of features such that they are independent (orthogonal) and are aligned with the directions of maximum variance. These linear combinations are known as principal components. The length of the vector built on the principal components is defined as the new metric. The use of PCA enabled us to maximize the information contained in each available metric while allowing us to formulate statistically sound independent variables based on these metrics. As an added benefit, it enabled us to reduce the number of models we needed to develop. Vector sum of the principal components in Eqs. (3) and (5) was deemed more appropriate for this analysis than linear sum because the latter is typically used in optimization use cases (e.g., simplex, gradient descent, etc.), whereas the former is usually preferred in predictive modeling, as it reflects the independence (orthogonality) of the principal components: we are only interested in the overall multidimensional change of the new variable.

Appendix B. COVID-19 metrics

Here we describe the metrics we used and derived for measuring different aspects of the COVID-19 pandemic.

B.1. Available COVID-19 metrics

Johns Hopkins University Coronavirus Research Center [8] provides data on the total daily number of confirmed COVID cases (cumulative and daily) and total number of deaths in the US at the county granularity level. The data, parameterized by county, look as shown in Figs. 9, 10, and 11:

The three curves tell the same story; Fig. 10 is a cumulative view of Fig. 9. In other words, e.g., by January 01, 2021, a total of 800K Los Angeles County residents have had COVID-19. The rise in cases in the first two figures matches the rise in the number of deaths in the third. They are strongly correlated, but not completely collinear (see Fig. 12).
Fig. 9. Covid new cases. Each point is a US county.

Fig. 10. Cumulative Covid confirmed cases. Each curve is a US county.

Fig. 11. Covid death toll. Each curve is a US county. The discontinuity in data for NYC is explained by Team, NYCDoH. and (dohmh) Covid-19 Response, M.H. and of Health, NYCD. and Team, MHDCR et al. [31].

B.2. Standard derived COVID metrics

We call COVID rate and COVID mortality standard derived metrics because they are computed and used in this research in the same way as, e.g., defined in [9].

B.2.1. COVID rate

COVID rate is an important measure of disease propagation in a county. It expresses the proportion of cases in the population. We computed COVID rate computed as \( \frac{N_c}{N_p} \), where \( N_c \) is the number of cases in the county, and \( N_p \) is the county’s population count. COVID rate characteristics are presented in Table 6.

B.2.2. COVID mortality

COVID mortality is an aggregate measure of severity of the disease and the state of preparedness, or lack thereof, of the county’s medical facilities, supplies, and personnel, including first responders and hospital workers, for handling the pandemic. To compute COVID mortality, we used the formula \( \eta = \frac{N_d}{N_c} \), where \( N_d \) is the day’s death toll, and \( N_c \) is the number of COVID cases. As described in Section 3.3, we time-shifted the \( N_d \) metric 4 days back to account for the average time between showing the symptoms and dying.

B.3. New (GRM) derived COVID metrics

To further our understanding of the diverse aspects of the pandemic represented by Johns Hopkins University’s Coronavirus Research Center (CRC)’s data, and the impact of mobility and socioeconomic metrics on the severity of the disease and people’s risk, we developed the following new metrics of COVID severity and risk.

B.3.1. Recovery time (average duration of being ill)

Disease recovery time is a very important metric for community planning and preparations for pandemics and other health-related disasters. It is an important indicator of disease severity and a critical element in estimating capacity and personnel requirements for medical treatment facilities and the impact of the disease on the economy. We used Eq. (2) in Section 3.5.2 for the average time to recover (duration of being ill). To derive Eq. (2), we used the SIR model [7]. In equilibrium, rate of recovery, multiplied by the duration, should be equal to the total concurrent number of recovered patients (Little’s Law, [32]):

\[
\frac{d(N_c - N_d)}{dt} \cdot T = N_c - N_d,
\]

(6)

where \( N_c \) is the number of cases; \( T \) is the average recovery time (duration of the person’s being ill), \( N_d \) is the number of deaths. Then \( d(N_c - N_d)/dt \) represents recovery speed, and \( N_c - N_d \) is the concurrent number of recovered COVID patients. While Little’s Law assumes

Table 6

| Covid rate |     |
|------------|-----|
| mean       | 2.40e−02 |
| std        | 2.94e−02 |
| min        | 9.96e−08 |
| max        | 2.06e−01 |
steady-state conditions (equilibrium), it is not a necessary condition, as Gilgur and Ferrandiz [33] demonstrated. This allows us to apply Eq. (6) even to situations of extreme overload of healthcare systems during 2020. To approximate equilibrium conditions, we used a small time interval $\Delta t$ over which to compute the left-hand side of Little’s Law equation ((6) (recovery rate), using the $\Delta(N_c - N_p)$ over the period of 1 day for each county as the approximation of the first derivative for that county. Recovery Time, Duration of Being Ill, as well as $\text{days, sick}$ have been used interchangeably throughout this article.

**B.3.2. Disease Severity Metric**

As we explained in Appendix B.1, Johns Hopkins University’s CRC provides three important standard derived metrics — total number of cases, new cases, and death toll. Each line represents a US county.

Fig. 13. Covid Disease Severity Metric — a linear combination of total COVID cases, new COVID cases, and death toll. Each line represents a US county.

**Algorithm 2 Computation of Disease Severity Metric**

1. Load data into a $N \times M$ matrix, where $M$ is the number of metrics, and $N$ is the number of rows;
2. Normalize the metrics using Eq. (7);
3. Perform PCA;
4. for pc in principal components do
   if pc < threshold then
     Discard pc;
   end if
5. end for
6. Compute the length of the eigenvector;
7. Save Data;

Here we assume that the intrinsic severity of COVID varies both within and between counties; therefore, we ran Algorithm 2 across all counties. The normalization is done as the standard min–max procedure:

$$X_{\text{normalized}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$ (7)

The formula used for computing the length of the eigenvector is:

$$|W_{\text{eigen}}(t)| = \sqrt{\sum_{i \in \text{PCS}} P_i(t)^2}$$ (8)

where $|W_{\text{eigen}}(t)|$ is the length of the eigenvector (returned by PCA) at time $t$; $P_i(t)$ are the principal components at time $t$; and PCS is the list of the principal components. The eigenvector calculated in (8) is the Disease Severity Metric (DSM) given by Eq. (3). Disease Severity Metric is a statistically sound composite metric combining the signal from new COVID cases, total COVID cases, and number of deaths. It carries as much information as its component metrics, but it removes the redundancy intrinsic in multicollinearity. For these reasons, we are using the Disease Severity Metric (DSM), rather than its elements, as an output variable of our model in this study.

**B.3.3. COVID risk**

Risk is the metric bringing together the severity of the disease and the probability of getting the disease. We defined the risk of COVID in Section 3.5.2 as given by Eq. (4). This was derived as follows:

$$R = D \times P_{\text{COVID}} \approx D \times \frac{N_c}{N_p} = D \times \kappa.$$ (9)

where $R$ is the risk; $P_{\text{COVID}}$ is the probability of contracting COVID; $D$ is severity measured by the Disease Severity Metric. $N_c$ is the number of confirmed cases, and $N_p$ is the population count. In turn, $\kappa = N_c/N_p$ is the variable defined as COVID rate.

**B.4. Correlations among COVID metrics**

Fig. 14 shows that: (i) Disease Severity Metric (DSM) is not correlated with any of the other COVID metrics. (ii) There is a weak positive correlation (0.4) between recovery time and COVID rate, $\kappa$. (iii) All other COVID metrics are not correlated with others.

This allows us to use these metrics, with the exception of Recovery Time ($\text{days, sick}$) and COVID Rate, as orthogonal (independent) aspects of the pandemic. Specifically, it eliminates the possibility of the Simpson’s paradox [21] and enabled us to properly formulate risk as a joint metric of two independent features — the likelihood of getting the disease and its severity (Section 3.5.2).

**Appendix C. Mobility metrics**

In Section 3.1, we stated our hypothesis that mobility, along with socioeconomic status, population density, and government measures, have an effect on the severity and propagation rate of the disease. Two mobility metrics have been made available by Facebook via the Data For Good Humanitarian Data Exchange (HDX) during the COVID-19 pandemic, (i) daily low-mobility fraction of the population and (ii) relative change in daily range of movement compared to pre-COVID days (February, 2020, with the exception of President’s day weekend (2/15–2/17/2020).

**C.1. Movement range change**

Range-of-Movement Change data must be taken with a grain of salt. The data available to us do not support year-over-year comparison. The scale of the vertical axis is the percentage of February, 2020 average range of movement data for the same county, with the exception of the President’s Day holiday weekend (see Fig. 15).

Yet we do see a significant drop from 5% wider travel in the first two weeks of March to a 32% reduction by early April, 2020, and even a drop to 38% in mid-April. This trend towards reduction was
offset by a steady increase, and by late May–early June the range of movement stabilized for a few months at nearly the February level, oscillating between \(-4\%\) and \(+6\%\). This was followed by a downward trend starting in September, 2020 and continuing through today, at about \(-10\%\) and going down. In movement-range data, it does seem that there was a weekly seasonality during late spring and all through the summer, albeit not very consistent. This may be explainable by the warm weather and people trying to get out more and farther away during the weekends. Notably, all major US holidays in 2020 (Easter — 4/12/2020, Veterans Day — 5/25/2020, Fourth of July — 7/04/2020, Labor Day — 9/07/2020, Thanksgiving — 11/26/2020, Christmas — 12/24/2020, and New Years — 01/01/2021) have been marked by a significant drop in the average range of movement in the US.

### C.2. Low-mobility fraction of the population

In addition to the range of daily movement, Facebook has made available the low-mobility fraction of each county’s population. It represents the proportion of county residents staying within a 600 \( \times \) 600 m square during the day (Bing Tiles [34]). Fig. 16 shows the average (across all 3300 US counties) pattern of the low-mobility fraction of the population.

We see that there is clearly a weekly seasonality, where the weekend fraction of low-mobility population increases by as much as 5 percentage points. We also see that, as COVID-19 started its spread in the United States, low-mobility fraction almost doubled to \(23\% - 36\%\) immediately in the first two months of the pandemic, reaching the peak during the week of April 12 and going down and finding a relatively stable level at \(18\% - 24\%\) during most of the spring and summer times of 2020, and starting to climb again in the days leading up to the US elections (November 7) and rising while increasing in the weekly amplitude during the following two months. A slight disturbance of the patterns occurred in July. In general, slight weekend increases on some of the major US holidays, usually considered family holidays, in 2020 (Easter — 4/12/2020, Thanksgiving — 11/26/2020, and Christmas — 12/24/2020) are noticeable. During COVID onset, the distribution varied by as much as 27 percentage points (\(17\% - 44\%\) on any single day). This range tightened during the year down to 14 points (\(18\% - 32\%\)). The overall tendency is to have a higher fraction of low-mobility population in high-density counties; however the outliers obfuscate this tendency.

### C.3. Mobility restraint index

We have two aggregate metrics of daily mobility during COVID (see Fig. 17).

The ranges of movement are measured as the number of Level 16 Bing tiles (600 \( \times \) 600 m squares). The “staying in place” refers to staying within the 600 \( \times \) 600 m square throughout the day. They are strongly negatively correlated, with the Pearson correlation of \(\rho = -0.94\) (Fig. 17). Yet they describe two different aspects of mobility restraint. To combine the information from these two metrics and eliminate redundancy, we formulate a metric of mobility restraint exercised by the population.

Like Disease Severity Metric (Appendix B.3.2), Principal Component Analysis (PCA) has been used to formulate this metric. Please see Appendix B.3.2 for details. In the case of Mobility Restraint Index, two principal components explain, respectively, 0.913 and 0.087 of the variance. By combining the PCA-transformed primary components into one mobility restraint metric, we have built a single variable efficiently describing both dimensions of mobility — Mobility Restraint Index, or MRI. The PCA algorithm 3 has been applied independently to each county. Following PCA, we defined MRI using Eq. (5) in Section 3.5.2.

**Algorithm 3 Computation of Mobility Restraint Index**

Load data into a NxM matrix, where M is the number of metrics, and N is the number of rows;

for county in counties do
  Normalize the two mobility metrics using Eq. (7);
  Perform PCA;
  for pc in principal components do
    if pc < threshold then
      Discard pc;
    end if
  end for
  Compute the length of the eigenvector;
end for
Concatenate all counties;
Save Data;

**Derivation** Similarly to DSM, we used the following formula for computing the length of the eigenvector:

\[
|V_{eigen}(t)| = \sqrt{P_0(t)^2 + P_1(t)^2},
\]  

where \(|V_{eigen}(t)|\) is the length of the eigenvector (returned by PCA) at time \(t\), \(P_0(t)\) is the first principal component at time \(t\), and \(P_1(t)\) is the second principal component at time \(t\). We call the eigenvector calculated in (10) the Mobility Restraint Index (MRI).

Fig. 18(a) shows the average MRI across all US counties. It has been computed independently for each of the approximately 3300 US counties.

### Appendix D. Demographic and socioeconomic metrics

**Purpose** In Section 3.1, we stated that mobility, socioeconomic status, population density, and government measures have an effect on the severity and propagation rate of the disease.
D.1. Population density

Population density is important for our study, because it impacts the frequency and duration of people’s interactions and therefore the rate of disease propagation, $\kappa$. Population density is measured as the number of people per square mile. Using latest US Census data for population and land area of each county, we computed density data as Eq. (11).

$$\rho_c = \frac{P_c}{A_c},$$  \hspace{1cm} (11)

where $\rho_c$ is population density, $P_c$ is the county’s population, and $A_c$ is its land area. Population density distribution is strongly right-skewed (Fig. 19). This complicates statistical analysis. To account for population density, we further broke down the numbers into 8 bins, using logarithmic scale, with bin 0 corresponding to the lowest-density counties and bin 7 to highest-density ones.

Having done that, we started to get meaningful insights into people’s mobility, disease severity, and risk during COVID.

D.2. Mean personal income

Of the variety of socioeconomic metrics, we chose each county’s mean personal income. We see that overall the personal income distribution has not changed significantly over the 3 years that we have collected the data for (2017–2019), Fig. 20. The bulk (non-outliers) distribution looks to be Gaussian, but that may be an artifact of averaging the income (Central Limit Theorem): personal income is not normally distributed. (Hartog 1980), (Fix 2017), (Kitov 2009).

D.3. Income inequality by county

The most common measure of income inequality is the Gini index. US Census Bureau defines Gini index as follows:
The Gini Index is a summary measure of income inequality. (…) The Gini coefficient ranges from 0, indicating perfect equality (where everyone receives an equal share), to 1, perfect inequality (where only one recipient or group of recipients receives all the income).

US Census Bureau provides an annual sampling of Gini Index data via American Community Survey (ACS). ACS covers 828 counties out of 3300. This is still sufficient to perform statistical analyses. Kitov [12] notes, “Effectively, the Gini coefficient has been almost constant during the last 60 years, as reported by the Census Bureau”. Thus, for the purposes of comparisons across counties, the mean is an adequate aggregate measure of personal income. Table 7 provides the summary of Gini Index for the US.

Appendix E. State government response measures

From the start of the pandemic, federal, state, and local governments have been taking measures to stem the spread of coronavirus. These measures have been of three types:

- Restrictions.
- Health and Safety support.
- Economic assistance.

Early in 2020, Oxford University came up with a way to quantify such measures, OxCGRT [35]. Initially, it was at the country aggregation level, but later they published such quantification for the US at the state level. We used these data to explore any correlations between government measures, socioeconomic factors, mobility, and vulnerability to COVID-19. Specifically, we used the following three indices from the OxCGRT data for the US states:

- for Restrictions: 'StringencyIndex',

Table 7

| Metric       | Gini Index |
|--------------|------------|
| mean         | 0.448458   |
| std          | 0.034617   |
| min          | 0.228800   |
| max          | 0.593200   |
Due to the absence of per-county OxCGRT data, we had to disregard county-level COVID response, which would have provided a better signal otherwise.

References

[1] Glodeanu A, Bilal U, Tosio PG. Social inequalities in mobility during and following the covid-19 associated lockdown of the madrid metropolitan area in spain. http://dx.doi.org/10.31235/ostf.io/np4e.

[2] Trias-Llimós S, Riffe T, Bilal U. Monitoring life expectancy levels during the covid-19 pandemic: Example of the unequal impact in spanish regions. http://dx.doi.org/10.1101/2020.06.03.2009792.

[3] Long J, Ren C. Socio-economic indicators correlate with daily mobility during the second-wave of the covid-19 pandemic. http://dx.doi.org/10.31235/ostf.io/tca34.

[4] Leung K, Wu JT, Leung GM. Real-time tracking and prediction of covid-19 infection using digital proxies of population mobility and mixing. Nature Commun 2021;12:1501. http://dx.doi.org/10.1038/s41467-021-21776-2.

[5] Porta M, Greenland S, Hernán M, dos Santos Silva I, Last J. A dictionary of epidemiology. Oxford quick reference. Oxford University Press; 2014, URL: https://books.google.com/books?id=4kZmAwAAQBAJ.

[6] Ridenhour B, Kowalik JM, Shay DK. Unraveling R0: Considerations for public health applications. Am J Public Health 2014;104:e32. http://dx.doi.org/10.2105/AJPH.2013.301704. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3935673/.

[7] Jackson MO. Social and economic networks. Princeton University Press; 2008, URL: http://www.jstor.org/stable/j.ctvc4gh1.

[8] Dong E, Du H, Gardner L. An interactive web-based dashboard to track covid-19 in real time. Lancet Infect Dis 2020;20:533–4. http://dx.doi.org/10.1016/s1473-3099(20)30120-1.

[9] Ritchie H, et al. 2021, URL: https://ourworldindata.org/mortality-risk-covid?country=USA.

[10] Schenkman S, Bouaquit A. From income inequality to social inequality: impact on health levels in an international efficiency comparison panel. BMC Public Health 2021;21:688. http://dx.doi.org/10.1186/s12889-021-10395-7.

[11] Jung E, Ro YS, Ryu HH, Shin SD, Moon S. Interaction effects between covid-19 outbreak and community income levels on excess mortality among patients visiting emergency departments. J Korean Med Sci 2021;36:e100. http://dx.doi.org/10.3346/jkms.2021.36.e100.

[12] Kitov I. Mechanical model of personal income distribution. SSRN Electron J http://dx.doi.org/10.2139/ssrn.1345790.

[13] Fix B. Evidence for a power theory of personal income distribution. http://dx.doi.org/10.31235/ostf.io/egeuu.

[14] de Oliveira Capela F, Ramirez-Marquez JE. Detecting urban identity perception via newspaper topic modeling. Cities 2019;93:72-83. http://dx.doi.org/10.1016/j.cities.2019.04.009.

[15] Carteni A, Di Francesco L, Martino M. How mobility habits influenced the spread of the covid-19 pandemic: Results from the italian case study. Sci Total Environ 2020;741:140489. http://dx.doi.org/10.1016/j.scitotenv.2020.140489.

[16] Banerjee S, Lian Y. Data driven covid-19 spread prediction based on mobility and mask mandate information. Appl Intell 2021. http://dx.doi.org/10.1007/s10489-021-02381-8.

[17] Noland RB. Mobility and the effective reproduction rate of covid-19. J Transp Health 2021;20:101016. http://dx.doi.org/10.1016/j.jth.2021.101016.

[18] Cruz AD. Forecasting philippine residential mobility percentage change during covid-19. SSRN Electron J 2020, URL: https://doi.org/10.2139/ssrn.3579317.

[19] Pak A, McBryde E, Adegbeye OA. Does high public trust amplify compliance with stringent covid-19 government health guidelines? A multi-country analysis using data from 102, 627 individuals. Risk Manag Healthc Policy 2021;14:293–302. http://dx.doi.org/10.2147/RMHP.S278774.

[20] Diaz Castro L, Cabello-Rangel H, Hoffman K. The impact of health policies and sociodemographic factors on doubling time of the covid-19 pandemic in mexico. Int J Environ Res Public Health 2021;18. http://dx.doi.org/10.3390/ijerph18052354.

[21] Malinowski J. Logic of simpson paradox. Logic Logical Philos 2005;14. http://dx.doi.org/10.12775/lhp.2005.013.

[22] Simpson JR. A combined biased-robust estimator for dealing with influence and collinearity in regression. 1993, http://dx.doi.org/10.21236/ada281793.

[23] Cooper I, Mondal A, Antonopoulos CG. A sir model assumption for the spread of covid-19 in different communities. Chaos Solitons Fractals 2020;139:110057. http://dx.doi.org/10.1016/j.chaos.2020.110057. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7321055/.

[24] Buckman SR, Glick R, Lanning KJ, Petrovsky-Nadeau N, Seitelman LM. Replicating and projecting the path of covid-19 with a model-implied reproduction number. 2020, URL: https://www.frbrf.org-economic-research/files/wp2020-24.pdf.

[25] Henry D, Ramirez-Marquez JE. Generic metrics and quantitative approaches for system resilience as a function of time. Reliab Eng Syst Saf 2012;99:114–22. http://dx.doi.org/10.1016/j.ress.2011.09.002.

[26] Gilgur A, Ramirez-Marquez J. Measuring community resilience: a bayesian approach. In: CESUN 2018. 2018, URL: https://www.slideshare.net/AlexGilgur1/measuring-community-resilience-a-bayesian-approach-cesun2018.

[27] Jewett RL, Mah SM, Howell N, Larsen MM. Social cohesion and community resilience during covid-19 and pandemics: A rapid scoring review to inform the united nations research roadmap for covid-19 recovery. Int J Health Serv: Plann Adm, Eval 2021;20731421997092. http://dx.doi.org/10.1177/0020731421997092.

[28] Kar B, Cochran DM. Risk communication and community resilience. Routledge; 2019, URL: https://play.google.com/store/books/details?id=6qiadwAAQBAJ.

[29] Gilgur A, Ramirez-Marquez J. Using machine learning to identify the factors of people’s mobility. In: INFORMS 2020 annual meeting. 2020, URL: https://bit.ly/3tvaDgh.

[30] Bennett J, Fry R, R K K. 2020, URL: https://www.pewresearch.org/fact-tank/2020/07/02/are-you-in-the-american-middle-class/.

[31] Teat, NYCDand (dohb) Covid-19 Response, Miland of Health, NYCDand Team, MHDCCR, Olson DR, Huynh M, et al. Preliminary estimate of excess mortality during the covid-19 outbreak — new york city, march 11–may 2, 2020. MMWR Morb Mortal Wkly Rep 2020;69:603–5. http://dx.doi.org/10.15585/mmwr.mm69053007.

[32] Hauser J, Urban G. From little’s law to marketing science. 2016, http://dx.doi.org/10.2147/RMHP.S278774.

Fig. 20. Personal Income Distribution during 2017, 2018, and 2019.
Gilgur A, Ferrandiz J. Little’s law assumptions: “but I still wanna use it!” the goldilocks solution to sizing the system for non-steady-state dynamics. MeasureIT; 2013, URL: http://www.cmg.org/measureit/issues/mit100/m_100_6.pdf.

DuVander A. Map scripting 101: an example-driven guide to building interactive maps with bing, yahoo!, and google maps. No Starch Press; 2010, URL: https://books.google.com/books/about/Map_Scripting_101.html?id=qAIh_2zNMw8C.

Guevarra E. oxcgrt: an interface to the oxford COVID-19 government response tracker API. 2020, r package version 0.1.0. URL: https://como-ph.github.io/oxcgrt/.

Alexander Gilgur is a Data Scientist and Systems Analyst with over 20 years of experience in a wide variety of domains – Control Systems, Chemical Industry, Aviation, Semiconductor Manufacturing, Information Technologies, Networking, Data Center Operations, and Social Sciences – and a solid track record of implementing his innovations in the industry. He has authored and co-authored a number of know-hows, publications, patents, and a technical textbook.

Alexander has teaching experience, including UC Berkeley’s School of Information (iSchool), where he taught Machine Learning in the MIDS program, Lincoln University, where he was invited on several occasions as a guest lecturer for MBA students, and a 4-hour course in Machine Learning for IT professionals, which he designed and delivered at the CMG Impact 2018 International Conference.

Currently Alexander works as a Research Data Scientist at Facebook and is a fifth-year Ph.D. Candidate in Systems Engineering and Management Sciences at Stevens Institute of Technology.

Jose Emmanuel Ramirez-Marquez is Associate Professor and Division Director of Enterprise Science and Engineering.

His research efforts focus on the development of mathematical models for the analysis and computation of system operational effectiveness — reliability and vulnerability analysis as the basis for designing system resilience. He also works at the intersection of evolutionary computation for the optimization of complex problems associated with system performance and design. His most recent research explores the interplay between data visualization and analytical decision making. In these areas, Dr. Ramirez-Marquez has conducted funded research for both private industry and government and, has published over 100 refereed manuscripts in technical journals, book chapters and industry reports.

Dr. Ramirez Marquez a former Fulbright Scholar, he holds degrees from Rutgers University in Industrial Engineering (Ph.D. and M.Sc.) and Statistics (M.Sc.) and from Universidad Nacional Autonoma de Mexico in Actuarial Science.