Linear Regularization-based Analysis and Prediction of Human Mobility in the U.S. during the COVID-19 Pandemic

Meghna Chakraborty (Corresponding Author)
Graduate Research Assistant
Department of Civil and Environmental Engineering
Michigan State University
428 S. Shaw Ln., East Lansing, MI 48824
Email: chakra43@msu.edu
ORCID: https://orcid.org/0000-0002-8369-1198

Md Shakir Mahmud
Graduate Research Assistant
Department of Civil and Environmental Engineering
Michigan State University
428 S. Shaw Ln., East Lansing, MI 48824
Email: mahmudmd@msu.edu
ORCID: https://orcid.org/0000-0003-3075-3196

Timothy J. Gates, Ph.D., P.E., PTOE
Associate Professor
Department of Civil and Environmental Engineering
Michigan State University
428 South Shaw Lane, East Lansing, MI 48824
Email: gatestim@msu.edu
ORCID: https://orcid.org/0000-0002-7429-0990

Subhrajit Sinha, Ph.D.
Postdoctoral Research Associate
Pacific Northwest National Laboratory
902 Battelle Blvd., Richland, WA 99354
Email: subhrajit.sinha@pnnl.gov
ORCID: https://orcid.org/0000-0003-4003-4590

Word Count: 6,225 words + 3 table(s) $\times$ 250 = 6,975 words

Submitted for presentation and publication: August 01, 2020

ABSTRACT

Since the increasing spread of COVID-19 in the U.S., with currently the highest number of confirmed cases and deaths in the world, most states in the nation have enforced travel restrictions resulting in drastic reductions in mobility and travel. However, the overall impact and long-term implications of this crisis to mobility still remain uncertain. To this end, this study develops an analytical framework that determines the most significant factors impacting human mobility and travel in the U.S. during the pandemic. In particular, we use Least Absolute Shrinkage and Selection Operator (LASSO) to identify the significant variables influencing human mobility and utilize linear regularization algorithms, including Ridge, LASSO, and Elastic Net modeling techniques to model and predict human mobility and travel. State-level data were obtained from various open-access sources for the period from January 1, 2020 to June 13, 2020. The entire data set was divided into a training data-set and a test data-set and the variables selected by LASSO were used to train four different models by ordinary linear regression, Ridge regression, LASSO and Elastic Net regression algorithms, using the training data-set. Finally, the prediction accuracy of the developed models was examined on the test data. The results indicate that among all models, the Ridge regression provides the most superior performance with the least error, while both LASSO and Elastic Net performed better than the ordinary linear model.

Keywords: COVID-19, mobility, number of daily trips per person, regularization algorithms, Ridge, LASSO, Elastic Net, linear regression

INTRODUCTION

The novel coronavirus (COVID-19) pandemic is delineating the global health crisis of our times and has had a tremendous impact on the way we understand our everyday lives as well as the world. Since its emergence in Asia late 2019, all continents in the world except Antarctica have been fighting the virus in earnest. The World Health Organization (WHO) (*1*) has declared COVID-19 a global pandemic on March 11, 2020 and the United States declared a national emergency on March 13, 2020 (*2*). As of July 15, 2020, more than 13.6 million cases of COVID-19 were confirmed in 215 countries around the globe (*3*). Among all countries, the U.S. has the highest number of confirmed cases and fatalities in the world due to COVID-19 (*3*). Several countries have closed their borders, exercised lockdowns, curfews, stay-at-home orders, and social distancing protocols, resulting in a sharp decrease in transport demand at local, regional, national, and also international levels. By March 24, 2020, more than 20 percent of the world's population has been ordered to remain at home as governments, health, and administrative organizations take extreme measures to protect their populations from the spread of the virus (*4*).

In the U.S., more than 40 states have already enforced stay-at-home order, the earliest being in California effective from March 19, 2020 (*5*). Only in the U.S., the mobility restrictions during the COVID-19 pandemic, in the form of travel bans, stay-at-home mandates, and lockdown policies, have impacted millions of people. Overall, human mobility has been severely impacted due to the travel bans and individual concerns to avoid public gatherings, resulting in tremendous economic impacts in transportation sectors. However, the overall influence and the long-term implications of this pandemic to mobility and transportation systems still remain uncertain at this point in time. Against the background of this unprecedented global crisis, questions remain as to how the different factors during the pandemic contribute to human mobility and travel.

With the increasing availability of high-quality data related to COVID-19, analyzing transportation and mobility during and after this crisis is absolutely imperative. Although there are still many unknowns, statistical models and analytical tools would help produce evidence-based research and policy interventions after COVID-19 outbreak. At the national level, Zhang et al. (2020) developed a COVID-19 impact analysis platform (*6*) that can inform users about the spread of COVID-19 in the U.S., and the effects of the virus spread and government orders on mobility and social distancing in the country using privacy-protected smartphone device location data, coupled with the census information. The platform gets updated daily and goes back to January 1, 2020, for benchmarking, and the results are scaled and aggregated to the entire population for both state and county levels (*7*). Gao et al. (2020) reported on the interactive web-based mapping platform (*8*) developed by the GeoDS Lab at the University of Wisconsin, Madison, with the support of the National Science Foundation RAPID program (*9*). This platform provides quantitative information on how people in different counties and states in the U.S. responded to the social distancing directives and guidelines. The platform integrates geographic information systems (GIS) and the daily updated human mobility patterns obtained from large-scale anonymous and aggregated smartphone data at the county-level (*10*).

With the availability of such large-scale data, this study applies an analytical framework that helps to understand how different factors impacted the daily trips of the U.S. population during the COVID-19 pandemic utilizing regularization regression techniques. Regression analysis is widely used in data analysis and machine learning (*11*), and is applied extensively in various applications, including transportation (*12, 13*). Among the various regression algorithms, linear regression is computationally one of the most efficient techniques and is often used as a starting

point for many different problems, although in many cases the linear regression model may be suboptimal. In the analysis of time-series data coming from a dynamical system, linear regression arises naturally in data-driven analysis of dynamical systems using the transfer operators, namely Perron-Frobenius and Koopman operators (*14–16*). More recently (*17, 18*) used robust optimization techniques to compute these operators from noisy data sets and the resulting optimization problem was a variation of the ordinary least squares (OLS), namely, least squares with regularization. Regularization is a standard technique used in data analysis to overcome some of the limitations of OLS, including the overfitting of training data and the susceptibility to noise in the data (*11*).

Literature on transportation analysis utilizing regularization is scant. Recently Polson and Sokolov (2017) developed a deep learning model utilizing a linear model that is fitted with ℓ_1 regularization to predict traffic flows. The study showed that the deep learning architecture was capable of capturing non-linear spatio-temporal effects in traffic and providing short-term prediction of traffic flow (*19*). Tan et al. (2011) proposed a semi-supervised Elastic Net regression method for pedestrian counting by utilizing sequential information between unlabelled samples and their temporally neighboring samples as a regularization term. The developed model was able to attain superior prediction performance and select representative features from the original set of features without losing their interpretability (*20*). Hasan et al. (2017) proposed statistical techniques to identify spatial relationships among road links in an urban road network to select predictors for a short-term traffic prediction model for a given road link. The study uses a time-lagged multiple linear regression method and utilizes two analytical methods, including the Elastic Net regularization and Granger Causality test using one year of traffic flow and speed data from the selected road network in Brisbane, Australia. For a given target link, the relevant predictors obtained by the Granger Causality and Elastic Net are used separately to build the respective traffic prediction models. The results show that Granger Causality-based traffic prediction model provides superior prediction accuracy than that using the Elastic Net regression (*21*). More recently, Battifarano and Qian (2019) explored the spatio-temporal correlations between the urban environment, traffic flow characteristics, and surge multipliers and proposed a general framework for predicting the short-term evolution of surge multipliers in real-time using a log-linear model with ℓ_1 regularization, integrated with pattern clustering. The modeling algorithm is validated by using Uber and Lyft data from Pittsburgh (*22*).

While there is a plethora of information available related to mobility and travel during this pandemic, it is absolutely critical to develop a robust methodological framework to accurately identify the key factors that influence human mobility during such health crisis, so that it may help direct the focus in policies and guidance in future. To this end, this study develops an analytical framework that helps determine the most significant factors affecting human mobility by utilizing linear regularization algorithms including the Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net modeling techniques.

DATA DESCRIPTION

Data for this analysis were collected and combined from multiple web-based open-access sources. The majority of data used in this analysis were requested and obtained from the COVID-19 Impact Analysis Platform developed at the Maryland Transportation Institute of the University of Maryland (UMD) (*6*). This platform provides both state and county-based information for 50 states in the U.S. and the District of Columbia. To match with the data available from other sources, for the

purpose of this study, state-wise data were requested from this platform. The relevant statewide data obtained from this source include the daily number of new COVID-19 confirmed cases per 1000 people, daily COVID-19 death rate, the number of daily trips per person, social distancing index, percent of out of county or state trips, transit mode share, population density, percentage of African American or Hispanic Americans, income and employment, unemployment rates, population staying at home, and percent working from home, among others. Social distancing index in the data set indicates the increasing space between individuals and decreasing frequency of contact and is represented as an integer from 0 to 100, where 0 indicates no social distancing in the state and 100 indicates all residents are staying at home.

Additional information was collected and appended with the data obtained from the UMD platform. Statewide percentage of the elderly population of different age categories, percentage of the population at different education levels, male to female ratio, and gross domestic product (GDP) information were collected from the U.S. Census Bureau (*23*) and joined with the principle data set. Information on the percentage of population staying at home during this time was obtained from the Bureau of Transportation Statistics (*24*). Moreover, the trends in mobility calculated from the number of trip directions requested by their user after comparing to the trip volume of January 13, 2020, reported by Apple, were further joined with the study data set (*25*). Also, data on community vulnerability index, which is a measure of poverty level estimated from seven indicators including no health insurance coverage, education level, income, gross rent as percentage of income, poverty, unemployment, and disability, were obtained from Surgo Foundation's open-access source (*26*). Furthermore, several states have exercised travel restrictions in the form of stay at home order, limitations on gatherings, domestic travel limitations, or school closures. These information was obtained from the COVID-19 State and Territory Action Tracker (*27*) provided by ESRI (Environmental Systems Research Institute) (*28*).

Following the joining of the data from various aforementioned sources, a thorough screening and quality check of the data was performed for any missing values. The final data set includes daily data starting from January 1, 2020 to June 13, 2020 from 50 U.S. states and the District of Columbia, and consists of a total of 8,415 observations for further analysis.

METHODOLOGY

In this section, we briefly discuss ordinary linear regression and its different regularized versions that are utilized in this study, including the Ridge, LASSO, and Elastic Net regression techniques.

Linear Regression

Consider the data set $\{y_i, x_{i1}, x_{i2}, \cdots, x_{iN}\}_{i=1}^{n}$ where y_i is the i^{th} observation of the dependent variable y and x_j, $j = 1, 2, \cdots, N$ are the N independent variables. In case of linear regression, the model tries to fit a straight line by minimizing the residuals. In particular, it assumes that the dependent variable can be expressed as a linear combination of the independent variables, as given by the following,

$$y_i = \alpha_0 + \alpha_1 x_{i1} + \alpha_2 x_{i2} + \cdots + \alpha_N x_{iN} + \varepsilon_i, \quad i = 1, 2, \cdots, n, \tag{1}$$

where ε_i is the residual.

In matrix form, the equation (1) can be written as

$$\mathbf{Y} = \mathbf{X}\alpha + \varepsilon, \tag{2}$$

where

$$\mathbf{Y} = [y_1, y_2, \cdots, y_n]^\top, \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1N} \\ 1 & x_{21} & \cdots & x_{2N} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nN} \end{pmatrix}, \alpha = [\alpha_0, \alpha_1, \cdots, \alpha_N]^\top, \varepsilon = [\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n]^\top.$$

The linear regression selects the parameters α_j's $(j = 1, \cdots, N)$ such that the norm of the residual for every y_i $(i = 1, \cdots, n)$ is minimized. Hence the optimal α is obtained as a solution of the following optimization problem,

$$\min_{\alpha} \parallel \mathbf{Y} - \mathbf{X}\alpha \parallel_2 . \tag{3}$$

where $\| \cdot \|_2$ is the 2-norm of a vector. The optimization problem (3) is convex and can be solved efficiently either using convex optimization techniques or analytically, such that the optimal $\alpha^\star$ is given by

$$\alpha^\star = \mathbf{Y}\mathbf{X}^\dagger, \tag{4}$$

where $\mathbf{X}^\dagger$ is the Moore-Penrose inverse of $\mathbf{X}$.

LASSO Regularization

Though linear regression is used extensively in many data-driven analyses and machine learning applications, it suffers from some drawbacks. One of the main disadvantages of the linear regression method is that it fails to identify the set of most important predictor variables. In many applications, it may be computationally intensive, and in some cases, even redundant, to include all explanatory variables in hand while fitting a model. Hence, it is advisable to determine the key predictors that have significant associations with the dependent variable and in such cases, the linear regression is not ideal (*29, 30*). To circumvent this challenge, prior studies (*29, 31*) suggested that the optimization problem (3) be modified as follows,

$$\begin{aligned} \min_{\alpha} \quad & \parallel \mathbf{Y} - \mathbf{X}\alpha \parallel_2 + \lambda_1 \parallel \alpha \parallel_1 \\ \text{subject to} \quad & \parallel \alpha \parallel_1 \leq t, \end{aligned} \tag{5}$$

where $\| \cdot \|_1$ is the 1-norm of a vector and the bound t is the tuning parameter. If t is large, it has no effect on the regression coefficients α_is and in this case, the solution to the optimization problem (5) approach the solution of normal linear regression optimization problem (3) in the limit of large t. However, when the bound t is small, the parameters α_is are constrained and hence are shrunk and are smaller versions of the original least squares estimates. The 1-norm minimization puts constraints on parameters that shrink coefficients towards zero. This is the shrinkage property of the LASSO regression that allows for a better interpretation of the model. By setting some of the α_is to zero, this technique identifies the most important variables associated with the dependent variable.

Ridge Regularization

Another major drawback of linear regression is that this algorithm has low bias and high variance (*11*). This means that the linear regression may perform well on the train data, but it may not generalize well to the test data set, thereby making the model performance unsatisfactory. In machine learning literature, this phenomenon is known as Bias-Variance trade-off (*11*). The intuition of bias-variance trade-off is explained in Figure 1(a). Usually, with a highly complex model, it is possible to fit the training data as closely as possible. In this case, the training error is zero

and the model is said to have a low bias. However, the highly complex model may not generalize well to the test data, thus making the test error large. This is due to the overfitting of the training data. The complex model, which overfits the training data and produces high test error, is said to have high variance. This situation is often reversed if the model considered is fairly simple. Ridge regression, which puts a 2-norm constraint on the set of coefficients, is able to efficiently overcome this challenge.

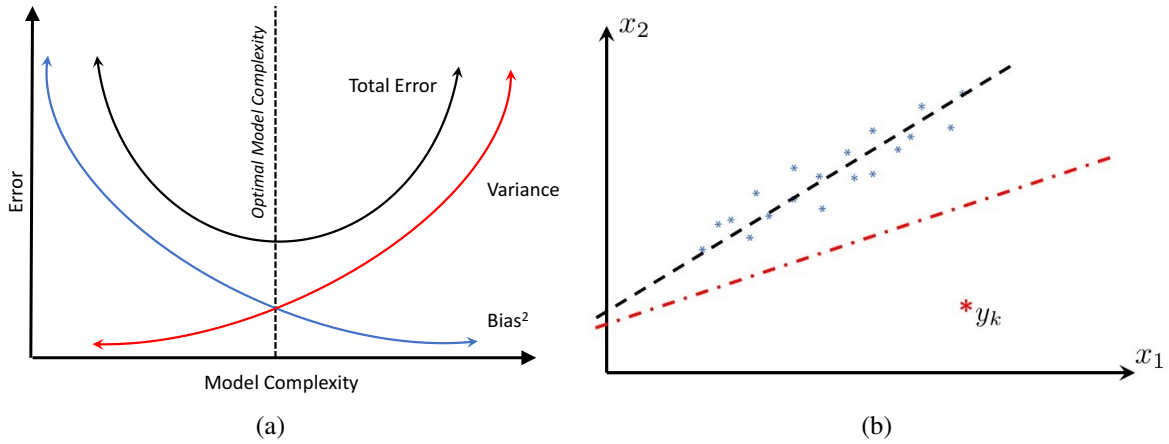


Figure 1 (a) Bias-variance trade-off. (b) Effect of outlier on linear regression.

Another drawback of ordinary linear regression is that the obtained model is highly influenced by the outliers in the training data set. For example, as in Figure 1(b), the outlier data point y_k results in the linear model fit as represented by the red line. However, it is obvious by looking at the overall data that the model fit depicted by the black line is the more appropriate linear fit to the data.

Additionally, on many occasions, the real-life data is noisy or uncertain. When the ordinary linear regression attempts to fit that noise in the data, it eventually results in overfitting and consequently degrading its performance for model prediction. As stated earlier, the data for this study were obtained mostly from smartphone devices, and the chances of acquiring this data may also be subject to individual user's discretion, so it is reasonable to assume that the data utilized in this study may contain some noise or uncertainty in it. To account for the noise in the data, it is assumed that there is some uncertainty, $\Delta\mathbf{Y}$ and $\Delta\mathbf{X}$, in the dependent and independent variables, respectively. It is assumed that the uncertainties in both $\mathbf{Y}$ and $\mathbf{X}$ are bounded, i.e. there exists some positive real number $\rho > 0$ such that $\| \Delta\mathbf{X} \|_2 \leq \rho$ and $\| \Delta\mathbf{Y} \|_2 \leq \rho$. With this, the optimization problem (3) is modified to a min-max optimization problem (*17, 18*) given by,

$$\min_{\alpha} \max_{\substack{\|\Delta\mathbf{X}\|_2 \leq \rho, \\ \|\Delta\mathbf{Y}\|_2 \leq \rho}} \| (\mathbf{Y} + \Delta\mathbf{Y}) - (\mathbf{X} + \Delta X)\alpha \|_2 . \tag{6}$$

Min-max optimization problems are generally hard to solve, but in this case, the optimization problem (6) can be equivalently expressed as a convex optimization problem as follows,

Theorem 1 *The optimization problem*

$$\min_{\alpha} \max_{\substack{\|\Delta \mathbf{X}\|_2 \leq \rho, \\ \|\Delta \mathbf{Y}\|_2 \leq \rho}} \| (\mathbf{Y} + \Delta \mathbf{Y}) - (\mathbf{X} + \Delta X)\alpha \|_2 . \quad (7)$$

is equivalent to the following,

$$\min_{\alpha} \| \mathbf{Y} - \mathbf{X}\alpha \|_2 + \lambda_2 \| \alpha \|_2^2, \quad (8)$$

where λ_2 is a positive real number, depending on the uncertainty bound ρ.

Proof. For proof, see (*17, 18*). □

The optimization problem (8), known as Ridge regression, is a convex problem and can be solved efficiently using any of the available convex optimization problem solvers. The parameter λ_2 is called the regularization parameter and it acts as a trade-off between the ordinary least squares cost and the cost on the coefficients α.

The LASSO and Ridge regression techniques are the modified linear regression algorithms that overcome the deficiencies of the linear regression algorithm and improves the performance of a linear model.

Elastic Net Regularization

Although the LASSO regression is efficient in feature selection from a large pool of variables, in case of high-dimensional data with few observations, the LASSO algorithm saturates (it selects up to *n* variables, where *n* is the number of observations) (*32*). Also, if some of the independent variables are highly correlated, the LASSO tends to select only one of those variables. The Elastic Net modeling (*32*), which is a combination of both the LASSO and Ridge regressions, is capable of overcoming this limitation. In particular, the Elastic Net uses the goods of both the worlds, the LASSO and Ridge regressions. The optimization problem for the Elastic Net can be expressed as,

$$\min_{\alpha} \| \mathbf{Y} - \mathbf{X}\alpha \|_2 + \lambda_1 \| \alpha \|_1 + \lambda_2 \| \alpha \|_2^2 . \quad (9)$$

The Ridge component of this technique, which corresponds to $\| \alpha \|_2^2$, makes the cost strongly convex and hence the optimization problem (9) has a unique minimum. But one problem with the Elastic Net modeling is the fact that often the optimization problem (9) is solved in two steps, which leads to higher bias and poor variance (*32*).

In this study, since the number of observations is much higher compared to the number of independent variables (dimension of the data), the LASSO regression is used to select the key variables. This is followed by analyzing the prediction capabilities of the various regression algorithms.

PRELIMINARY ANALYSIS

The preliminary analysis examines several explanatory variables of interest that were initially included in the analysis. The distribution of the dependent variable i.e., the number of daily trips per person during the analysis period (from January 1, 2020 to June 13, 2020) is shown in Figure Figure 2. As can be clearly seen from Figure 2, the number of daily trips remains consistent from January, 2020 to around mid-March, 2020. However, about the time when the U.S. started experiencing rapid community outbreaks of the virus and the country declared a national emergency, the number of daily trips drops substantially. Additionally, it is interesting to see that although the

number of COVID-19 cases per day continues to increase considerably over time, people started making more trips from around early May, 2020.

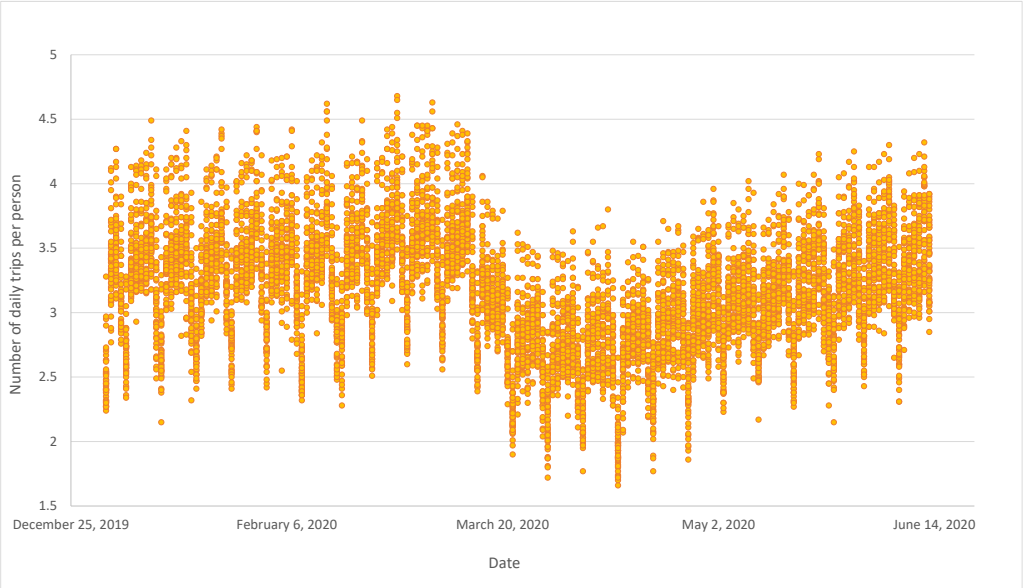


Figure 2 The number of daily trips per person during the analysis period.

As stated earlier, the compiled data set includes a large number of variables of interest that may impact the number of daily trips. Table 1 presents the descriptive statistics, including the minimums, maximums, means, and standard deviations of twenty-five variables that were initially considered as the set of potential explanatory predictors. For the purpose of scaling the data, population, median income, and GDP were included in their natural log forms.

TABLE 1 Descriptive Statistics of the Data

Parameter	Minimum	Maximum	Mean	Std. Dev
Number of daily trips per person	1.66	4.68	3.18	0.46
Social distancing index	10	83	32.62	16.38
Transit mode share	0.29	34.83	3.67	6.23
Percent of out of county trips	0.2	52.3	27.58	8.78
Number of new COVID-19 cases (per 1,000 people)	0	0.72	0.04	0.06
Total population	577,737	39,557,045	6,415,047	7,271,390
Percent of population older than 60 years	15	27	21.88	2.32
Median income (US dollars)	44,445	84,342	61,786.9	10,522.4
Percent of Hispanic or African American population	2.9	57	22.71	13.6
Unemployment rate	2.20	46.6	12.11	9.96
Percent of population working from home	2.30	55.7	16.04	11.62
Population staying at home	107,766	14,180,383	1,480,259	1,759,931
Gap between the first case and the stay-at-home order (days)	7	94	36.62	16.95
Socioeconomic status	0	1	0.51	0.29
State employee travel restriction	0	1	0.77	0.45
No stay-at-home order	0	1	0.25	0.44
Stay-at-home order	0	1	0.12	0.32
Stay-at-home guidance	0	1	0.63	0.48
School closures	0	1	0.96	0.19
Closure of some or all facilities	0	1	0.84	0.36
Recommended domestic travel limitation	0	1	0.26	0.44
Mandatory domestic travel limitation	0	1	0.2	0.40
Mandatory statewide mask policy	0	1	0.69	0.46
Driving mobility index	0	299.24	99.69	33.70
GDP in 2019 (10^9 U.S. dollars)	30.5	2,792.03	370.04	482.7
Percent of population aged 25 years or over with bachelor degree or higher	21.3	60.4	32.16	2.60

RESULTS

Model Selection

One of the major advantages of the LASSO regression is that it is able to efficiently select the explanatory variables that are important for predicting the dependent variable. In other words, from a set of N independent variables, the LASSO can suitably select $k < N$ independent variables. In particular, the LASSO solves the optimization problem,

$$\begin{aligned} \min_{\alpha} \quad & \| \mathbf{Y} - \mathbf{X}\alpha \|_2 + \lambda_1 \| \alpha \|_1 \\ \text{subject to} \quad & \| \alpha \|_1 \leq t \end{aligned} \tag{10}$$

and the k variables corresponding to the k largest absolute values of the coefficient vector α are chosen as the new reduced set of important independent variables.

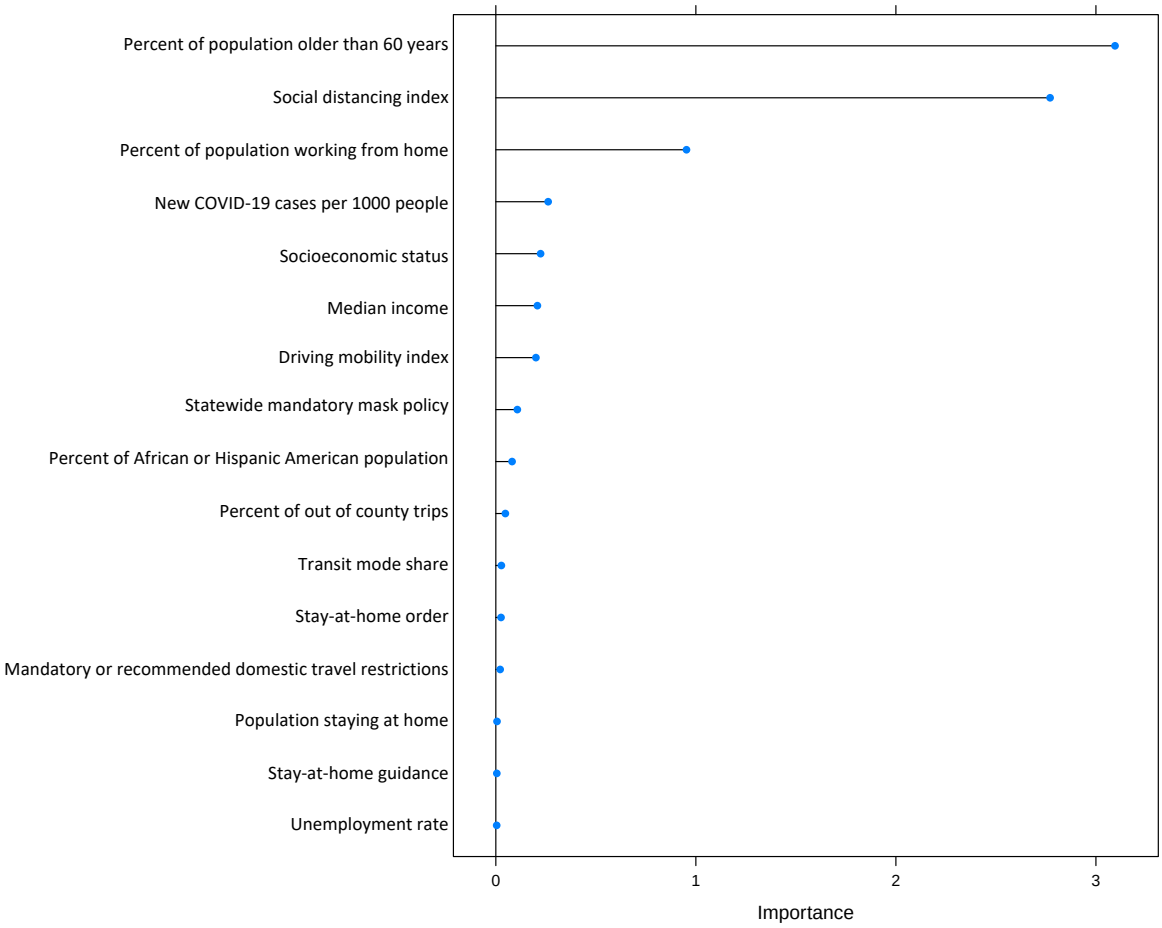


Figure 3 Important independent variables identified by the LASSO regression.

In this study, approximately 75 percent of the entire data were considered as training data and the remaining 25 percent of the data were selected for testing the model results. As stated before, the key variables were identified by the LASSO regression on the training data set with the regularization parameter varying from zero to one. Hence, the identification of the important variables was achieved by solving the optimization problem,

$$\begin{aligned} \min_{\alpha} \quad & \| \mathbf{Y} - \mathbf{X}\alpha \|_2 + \lambda_1 \| \alpha \|_1 \\ \text{subject to} \quad & 0 < \lambda_1 \leq 1. \end{aligned} \tag{11}$$

The optimization problem (11) was solved for each λ_1 with $\lambda_1 \in (0,1]$ and the optimal coefficients α were chosen for that particular λ_1 for which the cost function of the optimization problem (11) is the minimum. Ultimately, the LASSO regression problem (11) identified a total of sixteen predictors from the pool of twenty five features as the most important variables. The relative importance of the variables are shown in Figure 3.

Model Training

The final data set used in this study included information from January 1, 2020 to June 13, 2020. After splitting the data set into two parts for training and testing, the train and test data sets ranged from January 1, 2020 to May 3, 2020 and May 4, 2020 to June 13, 2020, respectively. The first (train) part was utilized for training the regression models, while the latter part (test) was used to test the efficiency of prediction from the obtained regression models. The sixteen key variables selected by the LASSO regression from the training data set were then trained for prediction by using all four modeling techniques, including the linear, Ridge, LASSO, and Elastic Net regression. This is accomplished by solving the optimization problems (3), (5), (8) and (9) to obtain the linear, LASSO, Ridge, and Elastic Net models, respectively. It is important to note that except for the linear model, the other optimization problems involve at least one regularization parameter (λ_1 and λ_2) and in all the optimization problems the λ_is were chosen such that, $0 < \lambda_i \leq 1$.

Interpretation and Discussion of the Model Results.

Table 2 compares the optimal coefficients of the final set of explanatory variables between all four regression modeling techniques. As can be seen from Table 2, the coefficients of the predictors are fairly comparable across all modeling techniques. As expected, when the number of new cases per 1000 people increases, the number of daily trips per person decreases. Similarly, with the increase in the social distancing index, the number of daily trips per person decreases. The percent of people aged more than 60 years and the number of people staying at home are negatively associated with the number of trips being made. Increasing concern regarding the higher health risk for the elderly population may have escalated the decrease in their number of daily trips.

Moreover, the number of trips per person decreases with a higher median income and socioeconomic status. Conversely, the number of daily trips per person increases with the increase in the unemployment rate. Additionally, the increase in transit mode share, driving mobility index, and percent of African and Hispanic American population are associated with the increase in the number of daily trips being made. Interestingly, when the stay-at-home order is implemented in a state, the count of daily trips per person goes lower. However, when the stay-at-home order is not stringently implemented and rather is recommended in a state, people are less likely to follow the guidance and tend to make more trips. Also, the number of trips drops when the states exercise mandatory or recommended domestic travel restrictions.

In situations where mandatory or recommended mask-wearing policy is in place, the number of trips increases, and this could partially be due to the fact that people may feel safer to go and travel when they themselves or people around them wear masks. It may also seem counter-intuitive that with the increase in the percent of the population working from home, the number of daily trips escalates. However, this can be explained as people may make more trips with the

flexibility in work schedules while working from home (WFH), compared to the usual scenario when a significant portion of the day is spent at office locations. Lastly, although the percent of out of county trips is found to be negatively associated with the number of daily trips being made in the linear, LASSO, and Elastic Net regression models, the expected positive association between these two variables is rightly captured in the Ridge regression technique. This is also confirmed by the error magnitudes, where the Ridge regression is providing the best model fit with the least root mean square error (RMSE) among all modeling techniques explored in this analysis. Overall, the most important factors to influence the number of daily trips include the number of new COVID-19 cases per 1000 people, social distancing index, percent of the population working from home, socioeconomic status, percent of the population older than 60 years, and driving mobility index, among others.

TABLE 2 Comparison of Model Coefficients between the Linear, Ridge, LASSO, and Elastic Net Regressions

Parameter	Linear	Ridge	LASSO	Elastic Net
Intercept	7.1495	4.6603	6.7524	6.9383
New COVID-19 cases per 1000 people	-0.2946	-0.2157	-0.2616	-0.2825
Social distancing index	-2.7815	-2.0815	-2.7709	-2.7543
Transit mode share	0.0284	0.0217	0.0277	0.0280
Percent of out of county trips	-0.0590	0.0918	-0.0476	-0.0504
Percent of people older than 60 years	-3.1940	-2.5281	-3.0955	-3.1422
Percent of Hispanic and African American population	0.0795	0.0441	0.0812	0.0791
Median income	-0.2416	-0.0522	-0.2078	-0.2244
Unemployment rate	0.0044	0.0024	0.0043	0.0043
Percent of population working from home	0.9885	0.3846	0.9531	0.9562
Population staying at home	-0.0064	-0.0075	-0.0059	-0.0063
Socioeconomic status	-0.2373	-0.1663	-.2238	-0.2307
Stay-at-home order	-0.0255	-0.0250	-0.0259	-0.0259
Stay-at-home guidance	0.0060	0.0027	0.0050	0.0055
Mandatory or recommended domestic travel restrictions	-0.0211	-0.0333	-0.0216	-0.0218
Mandatory statewide mask policy	0.1111	0.0873	0.1075	0.1093
Driving mobility index	0.2073	0.3174	0.2004	0.2097

Furthermore, the log-lambda plots shown in Figure 4 depicts how the independent variables

that enter in the model, vary across the Ridge, LASSO, and Elastic Net regression techniques as the regularization parameters change. When the regularization parameter λ is small, the contribution of the regularization part to the cost functions in the optimization problems (3), (5), (8) and (9) is small and as such all these optimization problems are reduced to the ordinary linear regression. However, as λ is increased, the weight on the regularization component in the optimization problems increases and as such, the coefficients of the independent variables become smaller and approach zero. It is clearly seen from Figure 4 that all the coefficients do not approach zero at the same time. In particular, the important variables remain non-zero for larger values of λ as compared to the relatively non-important variables. For example, in Figure 4(a), the variables including social distancing index, percentage of people older than 60 years, and percentage of people working from home remain the most important, because the coefficients of these variables remain non-zero as λ is increased. Similarly, from Figures 4(b) and (c), the important variables corresponding to the Ridge and Elastic Net regressions can be identified. However, based on the order of importance, the coefficients of the explanatory variables differ between the three regularization techniques. For example, as can be seen in Figure 4, in case of the LASSO and Elastic Net regressions, statewide mandatory mask policy is the third most important feature having a positive association with the dependent variable, whereas, in the Ridge regression, percent of out of county trips becomes the third in the order of importance among variables with a positive association with the number of daily trips. This difference is due to the characteristic of the Ridge regularization that reduces the norm of the coefficients more uniformly, while the LASSO model attempts to set as many coefficients to zero as possible.

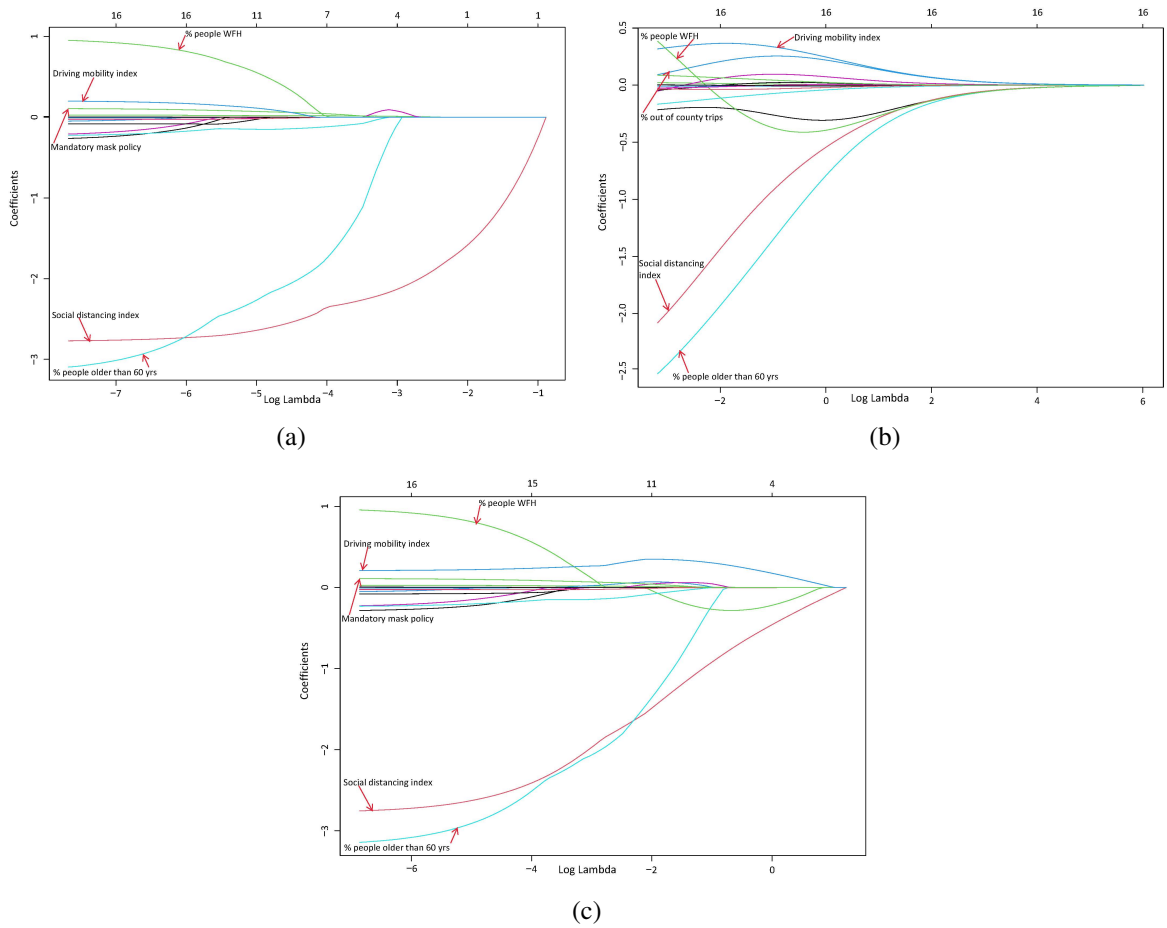


Figure 4 Relative importance of the independent variables in terms of log-lambda for (a) LASSO regression, (b) Ridge regression and (c) Elastic-net regression.

Prediction Performance

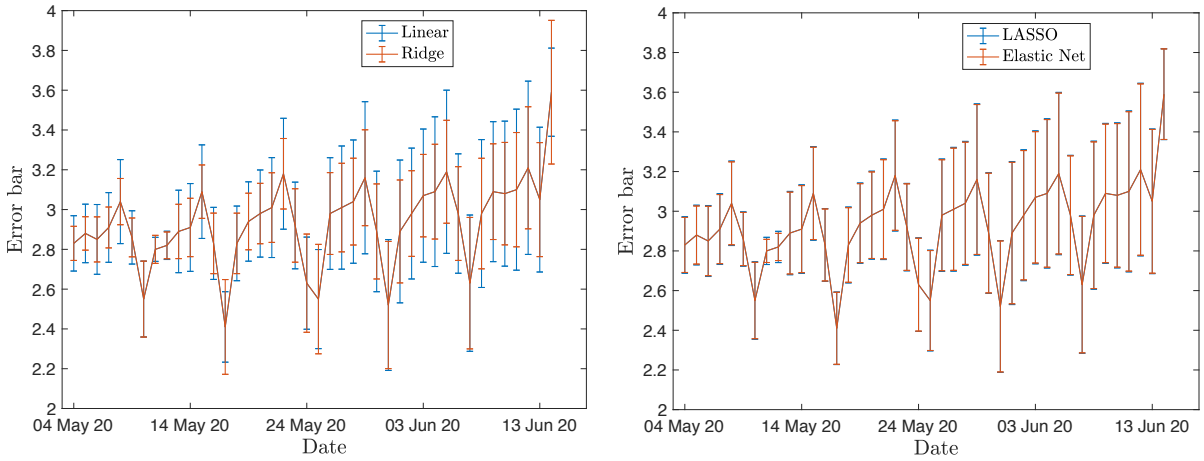
Following the development of the models using training data, the prediction of the dependent variable (the number of daily trips per person) is tested and compared between the four regression techniques. Essentially, the performance of the four different models on the test data set is evaluated, and the predicted values are compared with the observed values. Ultimately, the RMSEs from all models provide the measure of performance and efficiency of the models.

TABLE 3 Root mean square error (RMSE) results of training and test data

Model	Train	Test
Linear	0.65890	0.2774
LASSO	0.65806	0.2734
Elastic Net	0.65787	0.2748
Ridge	0.63612	0.2413

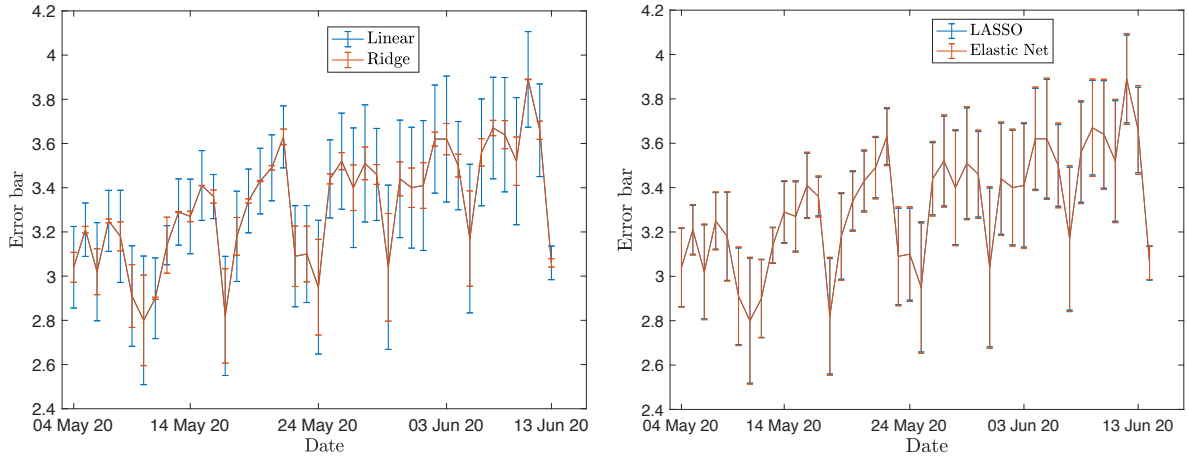
The RMSEs of the different models utilizing both the train and test data set are presented in Table 3. The comparison of the RMSEs between the four modeling techniques clearly shows that

the Ridge regression performs the best for both train and test data by having the least RMSE among all models. This is expected, as the Ridge regression provides superior prediction by overcoming the issue of overfitting with low variance and better generalization to the test data compared to other regularization methods (refer to Figure 1). Additionally, based on the RMSEs, both the LASSO and Elastic Net models provide better prediction performance compared to the ordinary linear regression.



(a) (b)

Figure 5 Error in prediction of daily trips per person in the state of California by (a) Linear and Ridge regression, (b) LASSO and Elastic Net.

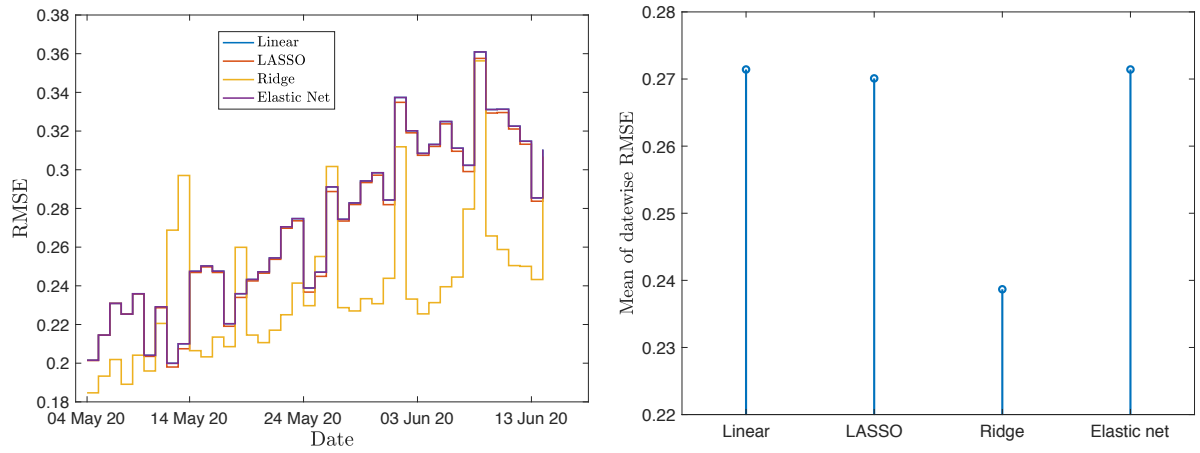


(a) (b)

Figure 6 Error in prediction of daily trips per person in the state of New York by (a) Linear and Ridge regression, (b) LASSO and Elastic Net.

For the purpose of showing how the different modeling techniques perform at an individual state level, graphical representations of RMSEs in the prediction of daily trips per person are given

separately for California and New York as examples in Figures 5 and 6, respectively. While the Figures 5(a) and 6(a) compare the errors in prediction by the ordinary linear regression and Ridge regression, the Figures 5(b) and 6(b) show the errors in prediction using the LASSO and Elastic Net modeling techniques. These figures clearly show that even at the individual state level, the Ridge model has the least error in prediction in both cases. Additionally, the LASSO and Elastic Net models perform almost similarly, which is expected because the RMSEs of the LASSO and Elastic Net over the entire test data set are found to be almost similar to each other (Table 3).



(a) (b)

Figure 7 (a) Root mean square error (RMSE) in prediction of the different models (Eq. (13)) over the analysis period. (b) Average date-wise root mean square error (RMSE).

Furthermore, the RMSEs in the prediction utilizing the four developed models over the analysis period are presented in Figure 7. Let y_{d_i} be the observed value of daily trips per person on the d^{th} day for the state i. Hence for the test data set, $d = 1,2,\cdots,41$ and $i = 1,2,\cdots,51$. Let $\hat{y}^p_{d_i}$ be the predicted daily trip per person on the d^{th} day for i^{th} state when using the p^{th} modeling technique. Here p is one of linear, Ridge, LASSO, or Elastic Net regression techniques. Therefore, the error in the prediction using the p^{th} model on the d^{th} day for i^{th} state is,

$$\varepsilon^p_{d_i} = y_{d_i} - \hat{y}^p_{d_i}. \tag{12}$$

The root mean square error in the prediction of the p^{th} model for the d^{th} day over all the states is,

$$r^p_d = \sqrt{\frac{\sum_{i=1}^{n_s}(\varepsilon^p_{d_i})^2}{n_s}} = \sqrt{\frac{\sum_{i=1}^{n_s}(y_{d_i} - \hat{y}^p_{d_i})^2}{n_s}}, \tag{13}$$

where $n_s = 51$ is the number of the states.

In Figure 7(a), r^p_d is plotted for all modeling techniques over the period ranging between May 4, 2020 to June 13, 2020 (test period). From Figure 7(a), it can once again be seen that the Ridge regression performs better for most of the test period, as its RMSE is usually lower compared to the other methods. The average of r^p_d across the test period, which can be expressed as,

$$r^p = \frac{\sum_{d=1}^{n_d} r_d^p}{n_d} = \left(\sum_{d=1}^{n_d} \sqrt{\frac{\sum_{i=1}^{n_s} (y_{d_i} - \hat{y}_{d_i}^p)^2}{n_s}} \right) \Big/ n_d, \tag{14}$$

where n_d is the number of days (equals to 41 in this study), is plotted in Figure 7(b). This plot also confirms that the Ridge regression model has the least error in prediction.

SUMMARY AND CONCLUSIONS

Since the emergence and rapid growth of novel coronavirus (COVID-19), countries worldwide are taking extreme measures to help prevent the spread of the virus. The U.S. is greatly hit by the pandemic and currently has the highest number of confirmed cases and deaths in the world due to COVID-19. Since the White House declared an emergency on March 13, 2020, most states in the U.S. implemented travel restrictions, stay-at-home orders, lockdowns, and social distancing protocols to combat the crisis, causing drastic reductions in travel demand at local, regional, and national levels. However, the overall impact and the long-term implications of this crisis to mobility still remain uncertain at this point in time. In order to understand these implications better, statistical models and analytical tools utilizing the increasingly available open-access data is the need of the hour. To this end, this study develops an analytical framework that helps to determine the most significant factors impacting human mobility and travel in the U.S. during the pandemic by utilizing linear regularization algorithms, including the Ridge, Least Absolute Shrinkage and Selection Operator (LASSO), and Elastic Net modeling techniques.

Data for this study were obtained for 50 states and the District of Columbia from various databases created and maintained to analyze the impacts of this pandemic. Relevant data on COVID-19 spread and its impact on mobility, social distancing, health, economy, and vulnerability of different groups were obtained from COVID-19 Impact Analysis Platform developed at the University of Maryland. Statewide percentages of the elderly population, population at different education levels, and GDP related information were obtained from the U.S. Census Bureau, while information on percentage of population staying at home was obtained from the Bureau of Transportation Statistics. Also, Data on travel restrictions, school closure, limitation on gatherings, stay at home order were obtained from COVID-19 State and Territory Action Tracker. Additionally, data were also collected from other sources on the mobility index and community vulnerability index. The compiled data set includes daily observations from 50 states and the District of Columbia starting from January 1, 2020 to June 13, 2020.

Evaluating an extensive data set requires advanced analysis techniques to identify the most important factors in explaining the response variable. Commensurate with analyzing such a rich data, this study employs different linear regularization techniques, including the Ridge, LASSO, and Elastic Net models, along with ordinary linear regression. The entire data set was split into two parts, where approximately 75 percent of the data (from January 1, 2020 to May 3, 2020) were used for training the models, and the remaining 25 percent (from May 4, 2020 to June 13, 2020) was used for testing the prediction performance. Determining the set of most important factors impacting the number of daily trips per person from the pool of several independent variables to increase the prediction accuracy was accomplished using the LASSO regression. The sixteen selected variables were further analyzed employing the linear, Ridge, Lasso, and Elastic Net regression techniques using the training data. Finally, the performance of the prediction is tested by feeding the test data into the models for all regression techniques.

The results of this study reveal that the coefficients of the predictors are fairly comparable across all modeling techniques. When factors including the number of new cases, social distancing index, median income, socioeconomic status, percent of people aged more than 60 years, the number of people staying at home, transit mode share, driving mobility index, and the percent of African and Hispanic American population increase, the number of daily trips per person decreases. Conversely, the number of daily trips per person increases with the increase in the unemployment rate and the percent of population working from home. Although, when the stay-at-home order is implemented, the number of daily trips drops, when the stay-at-home order is only recommended but not stringently enforced, people are less likely to follow the guidance and tend to make more trips. Also, the number of daily trips drops when the states exercise either mandatory or recommended domestic travel restrictions. The count of daily trips also increases with a mandatory or recommended mask-wearing policy in place. Lastly, although the percent of out of county trips is found to be negatively associated with the number of daily trips being made in the linear, LASSO, and Elastic Net regression models, the expected positive association between these two variables is correctly captured in the Ridge regression technique.

Furthermore, the developed models were used to predict the number of daily trips per person for all the states for a period of 41 days (from May 4, 2020 to June 13, 2020). Although all the developed models, namely ordinary linear, LASSO, Ridge and Elastic Net models, compare favorably, the Ridge regression model performed the best by having the least root mean square error (RMSE) in prediction among the different models. This result makes sense because the Ridge regression is robust in overcoming the issue of overfitting and thus generalizes better to the test data set, resulting in lesser prediction error. Also, other regularization techniques, including the LASSO and Elastic Net models, as well, performed superior to the ordinary linear regression.

The study is only the starting point to help understand the associations between different factors and human mobility during the COVID-19 pandemic. The authors of this study intend to expand this study to utilize county-based data to understand these associations from a more granular level. Moreover, it would be insightful to include additional variables into the analysis as potential independent variables. Also, additional factors such as vehicle miles traveled or the number of miles traveled per person would also be required to be explored in explaining mobility. From the modeling perspective, it is reasonable to argue that the available data is subjected to some uncertainties and future research should be carried out to explicitly take the uncertainties into account to derive at more precise models. Furthermore, as the crisis is moving on to the greater peaks in terms of the number of confirmed cases and deaths over time in the U.S., subsequent analysis is warranted with data from the following months (post June 13, 2020).

AUTHOR CONTRIBUTIONS

The authors confirm contribution to the paper as follows: study conception and design: Subhrajit Sinha, Meghna Chakraborty, Md Shakir Mahmud; data collection and preparation: Meghna Chakraborty, Subhrajit Sinha, Md Shakir Mahmud; analysis and interpretation of results: Subhrajit Sinha, Meghna Chakraborty, Md Shakir Mahmud; draft manuscript preparation: Meghna Chakraborty, Subhrajit Sinha, Md Shakir Mahmud, and Timothy Gates. All authors reviewed the results and approved the final version of the manuscript.

REFERENCES

1. *World Health Organization*, 2020 (accessed July 19, 2020).
2. *Proclamation on Declaring a National Emergency Concerning the Novel Coronavirus Disease (COVID-19) Outbreak*, 2020 (accessed July 19, 2020).
3. *COVID-19 Coronavirus Pandemic*, 2020 (accessed July 19, 2020).
4. Helen Davidson, *Around 20% of global population under coronavirus lockdown*, 2020 (accessed July 19, 2020).
5. COVID19.CA.GOV, *Stay at home Q&A*, 2020 (accessed July 19, 2020).
6. Maryland Transportation Institute, *University of Maryland COVID-19 Impact Analysis Platform*. University of Maryland, College Park, USA, 2020 (accessed July 19, 2020).
7. Zhang, L., S. Ghader, M. L. Pack, C. Xiong, A. Darzi, M. Yang, Q. Sun, A. Kabiri, and S. Hu, An interactive COVID-19 mobility impact and social distancing analysis platform. *medRxiv*, 2020.
8. Gao, Song and Rao, Jinmeng and Kang, Yuhao and Liang, Yunlei and Kruse, Jake, *Mapping Mobility Changes in Response to COVID-19*, 2020 (accessed July 19, 2020).
9. National Science Foundation, *National Science Foundation awards rapid response grants to support coronavirus (COVID-19) research*, 2020 (accessed July 19, 2020).
10. Gao, S., J. Rao, Y. Kang, Y. Liang, and J. Kruse, Mapping county-level mobility pattern changes in the United States in response to COVID-19. *SIGSPATIAL Special*, Vol. 12, No. 1, 2020, pp. 16–26.
11. Mitchell, T. M. et al., Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, Vol. 45, No. 37, 1997, pp. 870–877.
12. Chakraborty, M., S. Stapleton, M. Ghamami, and T. Gates, Safety effectiveness of All-Electronic Toll Collection systems. *Advances in Transportation Studies*, 2020.
13. Stapleton, S. Y., A. J. Ingle, M. Chakraborty, T. J. Gates, and P. T. Savolainen, Safety performance functions for rural two-lane county road segments. *Transportation research record*, Vol. 2672, No. 52, 2018, pp. 226–237.
14. Rowley, C. W., I. Mezic, S. Bagheri, P. Schlatter, D. HENNINGSON, et al., Spectral analysis of nonlinear flows. *Journal of fluid mechanics*, Vol. 641, No. 1, 2009, pp. 115–127.
15. Williams, M. O., I. G. Kevrekidis, and C. W. Rowley, A data–driven approximation of the koopman operator: Extending dynamic mode decomposition. *Journal of Nonlinear Science*, Vol. 25, No. 6, 2015, pp. 1307–1346.
16. Sinha, S., S. P. Nandanoori, and E. Yeung, Koopman Operator Methods for Global Phase Space Exploration of Equivariant Dynamical Systems. *arXiv preprint arXiv:2003.04870*, 2020.
17. Sinha, S., B. Huang, and U. Vaidya, Robust approximation of koopman operator and prediction in random dynamical systems. In *2018 Annual American Control Conference (ACC)*, IEEE, 2018, pp. 5491–5496.
18. Sinha, S., B. Huang, and U. Vaidya, On robust computation of koopman operator and prediction in random dynamical systems. *Journal of Nonlinear Science*, 2019, pp. 1–34.
19. Polson, N. G. and V. O. Sokolov, Deep learning for short-term traffic flow prediction. *Transportation Research Part C: Emerging Technologies*, Vol. 79, 2017, pp. 1–17.
20. Tan, B., J. Zhang, and L. Wang, Semi-supervised elastic net for pedestrian counting. *Pattern Recognition*, Vol. 44, No. 10-11, 2011, pp. 2297–2304.

21. Hasan, M. M., J. Kim, C. Prato, et al., Spatial variable selection methods for network-wide short-term traffic prediction. In *39 th Australasian Transport Research Forum (ATRF) Proceedings*, 2017.
22. Battifarano, M. and Z. S. Qian, Predicting real-time surge pricing of ride-sourcing companies. *Transportation Research Part C: Emerging Technologies*, Vol. 107, 2019, pp. 444–462.
23. *U.S. Census Bureau Data*, 2020 (accessed July 19, 2020).
24. *Bureau of Transportation Statistics, U.S. Department of Transportation*, 2020 (accessed July 19, 2020).
25. Apple, *Mobility Trends Reports*, 2020 (accessed July 19, 2020).
26. The COVID-19 Community Vulnerability Index (CCVI), *Community Vulnerability Index*, 2020 (accessed July 19, 2020).
27. National Governors Association, *COVID-19 State and Territory Actions Tracker*, 2020 (accessed July 19, 2020).
28. *Environmental Systems Research Institute*, 2020 (accessed July 19, 2020).
29. Tibshirani, R., Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, Vol. 58, No. 1, 1996, pp. 267–288.
30. Santosa, F. and W. W. Symes, Linear inversion of band-limited reflection seismograms. *SIAM Journal on Scientific and Statistical Computing*, Vol. 7, No. 4, 1986, pp. 1307–1330.
31. Bunea, F. et al., Consistent selection via the Lasso for high dimensional approximating regression models. In *Pushing the limits of contemporary statistics: contributions in honor of Jayanta K. Ghosh*, Institute of Mathematical Statistics, 2008, pp. 122–137.
32. Zou, H. and T. Hastie, Regularization and variable selection via the elastic net. *Journal of the royal statistical society: series B (statistical methodology)*, Vol. 67, No. 2, 2005, pp. 301–320.