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.
An analysis to identify the important variables for the spread of COVID-19 using numerical techniques and data science

Dee Pro F. Pasha, Alex Lundeen, Dilruba Yeasmin, M. Fayzul K. Pasha

ARTICLE INFO
Keywords:
COVID-19
Numerical techniques
Correlation
Least square method

ABSTRACT
Considering system theory, the socio-economic variables that constitute a society should be able to capture the system response such as the number of weekly COVID-19 cases. A numerical approach has been presented in this paper to answer two vital questions; which variables are more important and how many variables are needed to capture the dynamics of the spread. Using the theory of least squares regression, two types of problems have been set up and solved using multilinear regression (MLR) and nonlinear powered function known as NLR in this study. Numerical techniques were applied to pre- and post-process the data and the vast number of outputs. Total 43 socio-economic and meteorological variables from 31 counties in California in the United States resulted about 37.4 millions of combinations for the analysis. Results show that variables related to total population, household income, occupation, and transportation are more important than the others. The frequency of having higher correlation for a variable increases as more variables are combined with it. Similarly, correlation increases as the number of variables in a combination increases. Some 5-variable combinations can capture the dynamics of the spread with higher accuracy having correlation coefficient as high as 0.985.

1. Introduction

If society is considered to be a system, its variables should capture the system’s response and behaviors. There are many variables under different categories including demography, economy, culture, education, transportation, health, weather etc. that constitute a society. Some of these variables may be more important than the others in capturing system response and behavior. As the Coronavirus 2019 (COVID-19), also known as SARS-CoV-2 identified to be highly infectious [1,2], scientist and researchers are working to understand the dynamics of the spread of the virus using the predictive models in epidemics [3,4] and system behaviors to aid in the process of controlling the outbreaks [5].

According to the US Center for Disease Control [6], the primary method of COVID-19 spread is through respiratory droplets from human to human. Many researchers looked at the different factors that might affect the transmission of the droplets [7]. Some of those included demography [8,9], social connectedness and travel [10,11], economic, cultural, financial conditions [12,13], and meteorological characteristics [14–16]. Numerous relationships were found between these variables and the spread of COVID-19 via a variety of methods. Some of these relationships include higher COVID-19 susceptibility in older and more intergenerational populations and in lower temperature areas than in younger populations [8] and in higher temperature areas [14]. However, despite all of these predictions, the rate of infection continued to increase in summer affecting younger generations besides the elderly [17], disproving the rising average temperature hypothesis presented by many scientists.

While many factors that are possibly associated with COVID-19 cases have been identified discretely, a study is thus required where all the main socio-economic and meteorological variabilities are considered comprehensively. A systematic approach can include the social variabilities in a way that can quantify the impact of each of the variables and their combinations on the spread of COVID-19. IA comprehensive understanding of these combinations of variables and their magnitudes of influence on COVID-19 spread would greatly benefit the decision-makers to control the spread of COVID-19.

Data science and existing numerical techniques including single linear regression (SLR), multilinear regression (MLR), and nonlinear
regression (NLR) can be used to understand the relationships and their degrees of influence on COVID-19 spread. This paper outlines a numerical scheme to identify the important variables separately and combined and their impacts on the spread of COVID-19 using SLR, MLR, and NLR and well-known statistical assessment parameters. Please note that MLR with one independent variable is referred to as SLR in this study.

2. Methodology

Methodology consists of four steps. First step was data collection followed by data normalization and correlation analysis using MLR and NLR in the second and third steps respectively. In the final step, all the results were combined together to identify the important variables and their impacts on the spread of the virus. Literature search was conducted first to identify the variables that are postulated to be the reasons of spreading the virus. Since the number of weekly new COVID-19 cases, which is the dependent variable, was available throughout the study area, weekly time step was used as the temporal resolution. Hence, the meteorological variabilities were also converted into weekly basis. Different county data has been used to observe the impact of the data that are quasi-static in nature such as household income. Spatial resolution includes county level since different counties exhibit different weather and climatic behaviors but under similar state regulations.

To eliminate the bias in magnitude, data was normalized (standardized) using the following equation,

\[ Z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j} \]  

where, \( x_{ij}, \mu_j, \sigma_j, Z_{ij} \) were respectively any value, mean, standard deviation, and standardized values of variable \( j \).

Assuming linear relationship between observed independent variables \( X, (x_1, x_2, x_3, \ldots, x_n) \) and observed dependent variable vector \( Y (y) \) for single occurrence, the mathematical equation can be written as [18].

\[ Y = XA + \varepsilon \] (2)

\[ y = a_0 + a_1 x_1 + a_2 x_2 + a_3 x_3 + \ldots + a_n x_n + \varepsilon \] (3)

where, \( A (a_0, a_1, a_2, a_3, \ldots, a_n) \) is the regression coefficient vector and \( \varepsilon \) is the error vector. As mentioned, if a single independent (i.e., only \( x_1 \)) variable is used, the equation is referred to as SLR. If more than one independent variables are used, the equation is referred to as MLR in this paper. However, if the dependent variable is nonlinearly (using a power equation) related to the independent variable(s), the following equation which is known as NLR in this paper can be used.

\[ y = a_0 x_1^{a_1} x_2^{a_2} \ldots \ldots \ldots \ldots x_n^{a_n} + \varepsilon \] (4)

A logarithmic transformation of original data can be used to determine the coefficients of Eq. (4). Using the theory of least squares regression, the following equation can be used to solve for the regression coefficients \( A \) [18].

\[ A = (X'X)^{-1}X'Y \] (5)

where \( X \) is the matrix values of observed independent variables and \( t \) denotes the transpose of the matrix.
Two separate analyses one for MLR and the other for NLR were conducted to identify the important variables and their impacts in terms of frequency analysis and correlations. The following equation was used to calculate the correlation coefficient (CC).

\[
CC = \frac{n \sum x_i y_i - (\sum x_i)(\sum y_i)}{\sqrt{n \sum x_i^2 - (\sum x_i)^2} \sqrt{n \sum y_i^2 - (\sum y_i)^2}}
\]  

3. Application

State of California offers a wide range of variabilities in weather, social, and economic data under similar state regulation. Total 31 counties (Fig. 1) from different regions of the state were considered for the study. Total 43 independent variables have been selected under eight different categories; weather, demography, education, household income, occupation, health, transportation, and recreation. The definition, symbol, and source of data for each of these indicators are presented in Table 1. A brute force method was applied to calculate the total number of combinations assuming one (1) independent variable to five (5) independent variables combinations resulting 1,099,295 \((= 43C_1 + 43C_2 + 43C_3 + 43C_4 + 43C_5)\) combinations separately for MLR and NLR. These combinations were analyzed for each week separately for 17 weeks (from March 18 to July 15). Therefore, the combinations to analyze were 37,376,030 \((= 2 \times 17 \times 1,099,295)\). Considering the vastness of the combinations, MATLAB-R2018b [19] programming language was used to develop codes for computational purpose.

Frequency analysis includes identifying the number of weeks for which the correlation between a single independent and dependent variable was higher than a threshold value (CC \(>0.8\) in this case). Out of total 43, 19 variables were found highly correlated with the number of weekly cases (Fig. 2). The main categories of these variables are total population, household income, occupation, and transportation. Both MLR and NLR identify the same variables but with different magnitudes of correlation. For example, in household income category, VLI (i.e., very low income) has the highest impact and HI (i.e., high income) has the lowest impact on the spread of the virus. Similarly, in occupation category, SH (i.e., service in health) has the highest impact.

Some variables combined with the other variables may show better correlation. Therefore, considering the presence of each variable in a combination (2–5 variables combinations), the maximum possible number of occurrence for a single variable was calculated and used in the frequency analysis. Table 2 shows that an independent variable combined with other variable(s) can capture the dynamics better. As seen, the general trend of the impact is similar to one independent variable case.

![Fig. 2. Frequency of independent variable affecting the number of cases with CC greater than 0.80 for MLR and NLR with single independent variable.](https://map.google.com)
Variables with minimum or no impact in single variable case can have more impact when combined with other variables. Considering the vastness of the combinations, average $CC$ under each number of combination for all 17 weeks of results was used to assess the impact of the number of variables in a combination. The average $CC$ represents the trend and average impact of the number of variables on the dynamics of the spread (Fig. 3). As seen, $CC$ increases with the increase of number of variables in a combination. This is expected, since one or few variable(s) may not be enough to represent the true dynamics of the COVID-19 spread. The slopes of the curve are found steeper in the beginning than towards the end representing smaller contributions by additional variables. Similar to Fig. 2 and Table 2, the NLR is capable of assessing the impacts of each combinations separately causing the average $CC$ smaller than MLR.

Another analysis has been conducted to observe how average $CC$ varies by week. It is found that the average $CC$ varies from 0.50 to 0.96.

### Table 2

| Independent Variable | MLR 2 Var | MLR 3 Var | MLR 4 Var | MLR 5 Var | NLR 2 Var | NLR 3 Var | NLR 4 Var | NLR 5 Var |
|----------------------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|-----------|
| AT                   | 44%       | 70%       | 84%       | 92%       | 32%       | 59%       | 78%       | 89%       |
| McF                  | 44%       | 70%       | 84%       | 92%       | 27%       | 53%       | 73%       | 86%       |
| MnT                  | 44%       | 70%       | 84%       | 92%       | 33%       | 61%       | 80%       | 91%       |
| Rh                   | 44%       | 70%       | 84%       | 92%       | 29%       | 56%       | 76%       | 88%       |
| Sr                   | 44%       | 70%       | 85%       | 92%       | 22%       | 48%       | 70%       | 85%       |
| Ws                   | 44%       | 70%       | 84%       | 92%       | 21%       | 47%       | 69%       | 84%       |
| TP                   | 100%      | 100%      | 100%      | 100%      | 83%       | 92%       | 96%       | 99%       |
| PD                   | 45%       | 71%       | 85%       | 92%       | 24%       | 51%       | 72%       | 86%       |
| A18                  | 45%       | 71%       | 85%       | 92%       | 34%       | 60%       | 79%       | 89%       |
| A65                  | 44%       | 69%       | 84%       | 92%       | 36%       | 63%       | 81%       | 91%       |
| AO                   | 45%       | 70%       | 85%       | 92%       | 22%       | 48%       | 70%       | 85%       |
| H                    | 45%       | 70%       | 85%       | 92%       | 39%       | 65%       | 82%       | 92%       |
| WNH                  | 43%       | 69%       | 84%       | 92%       | 39%       | 66%       | 83%       | 92%       |
| W                   | 43%       | 70%       | 85%       | 92%       | 21%       | 48%       | 70%       | 85%       |
| B                   | 43%       | 69%       | 84%       | 92%       | 25%       | 51%       | 72%       | 86%       |
| A                   | 45%       | 71%       | 85%       | 92%       | 25%       | 52%       | 74%       | 87%       |
| Orb                  | 43%       | 70%       | 85%       | 92%       | 32%       | 62%       | 82%       | 92%       |
| HS                   | 44%       | 70%       | 85%       | 92%       | 39%       | 65%       | 81%       | 91%       |
| BS                   | 45%       | 71%       | 85%       | 92%       | 36%       | 61%       | 79%       | 90%       |
| VLI                  | 100%      | 100%      | 100%      | 100%      | 77%       | 86%       | 92%       | 95%       |
| Li                   | 100%      | 100%      | 100%      | 100%      | 68%       | 80%       | 89%       | 94%       |
| LML                  | 100%      | 100%      | 100%      | 100%      | 61%       | 76%       | 88%       | 94%       |
| MI                   | 100%      | 100%      | 100%      | 100%      | 73%       | 84%       | 91%       | 95%       |
| HI                   | 100%      | 100%      | 100%      | 100%      | 59%       | 78%       | 89%       | 94%       |
| TO                   | 100%      | 100%      | 100%      | 100%      | 72%       | 85%       | 92%       | 96%       |
| MBIA                 | 100%      | 100%      | 100%      | 100%      | 62%       | 79%       | 90%       | 95%       |
| Serv                 | 100%      | 100%      | 100%      | 100%      | 71%       | 84%       | 91%       | 95%       |
| SO                   | 100%      | 100%      | 100%      | 100%      | 66%       | 79%       | 88%       | 94%       |
| NRCM                 | 100%      | 100%      | 100%      | 100%      | 66%       | 81%       | 91%       | 95%       |
| PTM                  | 100%      | 100%      | 100%      | 100%      | 70%       | 80%       | 89%       | 93%       |
| SH                   | 100%      | 100%      | 100%      | 100%      | 80%       | 90%       | 96%       | 99%       |
| Dia                  | 44%       | 70%       | 84%       | 92%       | 31%       | 57%       | 75%       | 87%       |
| Ob                   | 44%       | 70%       | 84%       | 92%       | 26%       | 52%       | 72%       | 85%       |
| CA                   | 100%      | 100%      | 100%      | 100%      | 71%       | 83%       | 90%       | 95%       |
| CC                   | 100%      | 100%      | 100%      | 100%      | 64%       | 78%       | 88%       | 94%       |
| PT                   | 69%       | 90%       | 97%       | 99%       | 45%       | 70%       | 87%       | 94%       |
| Ws                   | 100%      | 100%      | 100%      | 100%      | 35%       | 61%       | 80%       | 90%       |
| TC                   | 100%      | 100%      | 100%      | 100%      | 30%       | 55%       | 75%       | 87%       |
| WH                   | 100%      | 100%      | 100%      | 100%      | 57%       | 76%       | 88%       | 93%       |
| SAD                  | 45%       | 71%       | 85%       | 93%       | 22%       | 48%       | 70%       | 84%       |
| SAI                  | 44%       | 70%       | 85%       | 93%       | 22%       | 47%       | 70%       | 84%       |
| CH                   | 44%       | 70%       | 84%       | 92%       | 23%       | 49%       | 70%       | 84%       |
| Mu                   | 44%       | 71%       | 85%       | 93%       | 26%       | 54%       | 75%       | 88%       |

(Dig. 2). However, variables with minimum or no impact in single variable case can have more impact when combined with other variables.

Considering the vastness of the combinations, average $CC$ under each number of combination for all 17 weeks of results was used to assess the impact of the number of variables in a combination. The average $CC$ represents the trend and average impact of the number of variables on the dynamics of the spread (Fig. 3). As seen, $CC$ increases with the increase of number of variables in a combination. This is expected, since one or few variable(s) may not be enough to represent the true dynamics of the COVID-19 spread. The slopes of the curve are found steeper in the beginning than towards the end representing smaller contributions by additional variables. Similar to Fig. 2 and Table 2, the NLR is capable of assessing the impacts of each combinations separately causing the average $CC$ smaller than MLR.

Another analysis has been conducted to observe how average $CC$ varies by week. It is found that the average $CC$ varies from 0.50 to 0.96.
for 1- to 5- variable combinations for both MLR and NLR. The average CC for a particular combination fluctuates more in NLR than in MLR. The highest and the lowest CC are found for Week 2 (March 25 – March 31) and Week 11 (May 27 – June 2) respectively.

One of the highest CC value combinations (TP, VLI, SH, PT, and CH) has been used to observe the dynamics and predictability of the regression coefficients for Week 3 (Fig. 4). As seen, both MLR and NLR can generally capture the pattern of the observed cases for all the counties. While MLR slightly over and under predicts, the NLR performs better prediction for the lower number of cases.

4. Conclusion

A numerical approach, which consists of the applications of MLR and NLR has been presented to observe the correlation between different socio-economic and meteorological variables and the weekly number of COVID-19 cases. Results show that 8 independent variables under total population, household income, occupation, and transportation categories are highly correlated with COVID-19 cases. As the given number of independent variables in a combination increases, the frequency of occurrences for higher correlation coefficient also increases. Similarly, the average CC, which is calculated considering all weeks results increases with the number of independent variables in a combination. Some highly correlated combinations can capture and predict the system behavior with high accuracy. The correlation coefficients for these combinations can be as high as 0.985. These observations can be used to develop a predictive model.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] WHO World Health Organization, World Health Organization Media Briefing. Director General’s Opening Remarks at the Media Briefing on COVID-19 - 11 March 2020, 2020. Retrieved online on June 30, 2020 at, https://www.who.int/dg/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19—11-march-2020.
[2] JHU Johns Hopkins University, ‘COVID-19 Dashboard by the Center for Systems Science and Engineering (CSSE)’, Coronavirus Resource Center, 2020. Retrieved online on July 11, 2020 at, https://coronavirus.jhu.edu/map.html.
[3] W.O. Kermack, A.G. McKendrick, Contributions to the mathematical theory of epidemics. II - the problem of endemicity, Proc. Roy. Soc. Lond. A 138 (834) (1932) 55–83, 1932.
[4] D. Fanelli, F. Piazza, Analysis and forecast of COVID-19 spreading in China, Italy and France, Chaos, Solit. Fractals 134 (2020) 109761, https://doi.org/10.1016/j.chaos.2020.109761.
[5] K.R. Wigginton, A.B. Boehm, Environmental engineers and scientists have important roles to play in stemming outbreaks and pandemics caused by enveloped viruses, Environ. Sci. Technol. 54 (2020) 3736–3739.
[6] CDC Center for Disease Control, Novel Coronavirus Factsheet - what You Should Know about COVID-19 to Protect Yourself and Others, 2020. Retrieved on July 15 at, https://www.cdc.gov/coronavirus/2019-ncov/downloads/2019-ncov-factsheet.pdf.
[7] L. Dietz, P.F. Horve, D.A. Coll, M. Fretz, J.A. Ellen, K.V.D. Wymelenberg, 2019 novel Coronavirus (COVID-19) pandemic: built environment considerations to reduce transmission, Applied and Environmental Science: American Society for Microbiology 5 (2) (2020) 1–13 (March/April 2020).
[8] J.B. Dowd, L. Andriano, D.M. Brael, V. Rotondi, P. Block, X. Ding, Y. Liu, M.C. Mills, Demographic science aids in understanding the spread and fatality rates of COVID-19, PANS 117 (No. 18) (2020) 9096–9096.
[9] M. Ahmadi, A. Sharafi, S. Dorosti, S.J. Ghoushchi, N. Ghanbari, Investigation of effective climatology parameters on COVID-19 outbreak in Iran, Sci. Total Environ. 729 (2020) (2020) 138705. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7293739/
[10] T. Kuchler, D. Ruszel, J. Streobel, The Geographic Spread of COVID-19 Correlates with Structure of Social Networks as Measured by Facebook, 2020. Retrieved on April 07, 2020 from, https://arxiv.org/abs/2004.03055.
[11] S.H. Ebrahim, Q.A. Ahmed, E. Gozzer, P. Schlegenhof, Z.A. Memish, COVID-19 and community mitigation strategies in a pandemic, BMJ 2020, https://doi.org/10.1136/bmj.m1066, 368:m1066, https://www.bmj.com/content/368/bmj.m1066.full.pdf.
[12] R. Mogil, J. Spjiker, The influence of social and economic ties to the spread of COVID-19 in europe, SOC AIRXIV Papers (2020), https://doi.org/10.31235/osf.io/sbfm., https://osf.io/preprints/socarxiv/sb8xn/.
[13] W. Koma, S. Artiga, T. Neuman, G. Claxton, M. Rae, J. Kates, J. Michaud, Low-Income and Communities of Color at Higher Risk of Serious Illness if Infected with Coronavirus, 2020. Retrieved on May 30, 2020 at, https://www.kff.org/coronavirusCOVID-19-issue-brief/low-income-and-communities-of-color-at-higher-risk-of-serious-illness-if-infected-with-corona/.
[14] J. Liu, J. Zhou, J. Yao, X. Zhang, L. Li, X. Xu, B. Wang, S. Fu, T. Niu, J. Yan, Y. Shi, X. Ren, J. Niu, W. Zhu, S. Li, B. Luo, K. Zhang, Impact of meteorological factors on the COVID-19 transmission: a multi-city study in China, Sci. Total Environ. 726 (2020) (2020) 138513, https://doi.org/10.1016/j.scitotenv.2020.138513.
[15] S.I. Priyadarshini, M. Suresh, Factors influencing the epidemiological characteristics of pandemic COVID-19: a TISM approach, Int. J. Healthc. Manag. (2020), https://doi.org/10.1080/20479700.2020.1755804.
[16] M.M. Sajadi, P. Habibzadeh, A. Vintzileos, S. Shokouhi, F. Miralles-Wilhelm, A. Anthony, Temperature, Humidity, and Latitude Analysis to Predict Potential Spread and Seasonality for COVID-19, 2020. Research Paper retrieved on July 10, 2020 from, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3550308.
[17] L.I. Maragakis, Coronavirus and COVID-19: Younger Adults Are at Risk, Too, Johns Hopkins Medicine Publication, 2020. Retrieved on June 25, 2020 from, https://www.hopkinsmedicine.org/health/conditions-and-diseases/coronavirus/corona-virus-and-covid-19-younger-adults-are-at-risk-too.
[18] S. Chapra, R.P. Canale, Numerical Methods for Engineers, seventh ed., 2015, 2015. ISBN-13: 978-0073397924.
[19] MATLAB, Version R2018b, The MathWorks Inc, 2019. https://www.mathworks.com/.