Understanding the Effects of Human Factors on the Spread of COVID-19 using a Neural Network

Anirudh Chhabra, Dhruv Patel, Xin Li, Lynn Pickering,
Javier Viaña and Kelly Cohen

February 27, 2021

2020 7th International Conference on Soft Computing & Machine Intelligence (ISCMI)
Stockholm, Sweden, 2020, pp. 121-125. (Preprint and Prereview Version)

Understanding the Effects of Human Factors on the Spread of COVID-19 using a Neural Network

Anirudh Chhabra[*], Dhruv Patel[†], Xin Li[‡], Lynn Pickering[§], Javier Viaña[¶], and Kelly Cohen[‖]
University of Cincinnati
Cincinnati, Ohio 45221, U.S.A.
Email: [*]chhabrad@mail.uc.edu, [†]patel4db@mail.uc.edu, [¶]KratosOmega@iCloud.com, [§]pickerln@mail.uc.edu,
[‡]vianajr@mail.uc.edu, [‖]cohenky@ucmail.uc.edu

***Abstract*—During the spread of an infectious disease such as COVID-19, the identification of human factors that affect the spread is a really important area of research. These factors directly impact the spread of such a disease and are important in identifying the various regions that are at a higher risk than others. This allows for an optimal distribution of resources according to predicted demand. Models such as the SIR framework exist and are very good at representing the spread of diseases and can incorporate multiple factors that resemble real-life scenarios. The primary issue in this area is the identification of relevant variables. In this study, a residual analysis is presented to downsize the dataset available and shortlist the small number of variables classified as absolutely necessary for disease modeling. The performance of different datasets is evaluated using an Artificial Neural Network and regression analysis. The results show that the drop in performance is reasonable and this approach can be automated in the future as it offers a small dataset containing a few variables against a large dataset with possibly hundreds of variables.**

***Index Terms*—COVID - 19, Infectious Diseases, Neural Network, Artificial Intelligence, Variable Reduction, Human Factors**

I. Introduction

COVID - 19 has proven to be the greatest challenge for humanity in recent times. To protect their citizens and prevent the overburdening of resources such as medical facilities and management drugs during peak virus transmission periods, governments around the world enforced quarantine measures. Due to the negative effects of such measures on the economy, these measures can only be applicable for short periods of time [1]. As countries look to relax these guidelines, prediction of virus hotspots is of extreme importance as this allows authorities to impose strict measures only where it is required. This can have positive effects on the economy as the world looks to recover from months of extreme economic stress. To properly predict such hotspots, or regions at high risk, understanding the human factors affecting the situation is of great importance.

Human factors play a crucial role in the spread of infectious diseases such as COVID-19. These factors basically translate to the number of people a person comes in contact with. Such information can generally explain where there might be a faster spread of the disease. If a relationship can be quantified between such factors and the spread of the disease, it allows the authorities to identify regions potentially at a higher risk than others and allow for a proper allocation of resources. As we have seen in the past few months, optimal allocation of resources is highly desirable. Furthermore, having knowledge and confidence about disease spread predictions will help authorities to fight situations like these more effectively.

Infectious disease modeling dates back to the early 20th century. Simple compartmental models like the SIR model [2] were developed to understand the spread of infectious diseases and remain an important category of techniques to understand infectious disease modeling to this day. Multiple modifications have been made to the SIR model such as the SEIR model which introduces another category known as ”Exposed Population” [3]. These tend to explain the dynamics of disease modelling in a more intuitive and logical manner. Recent papers have incorporated policies such as social distancing, economic trade-offs, age, fatality rates in older populations, and other policies into the compartmental model framework (See Refs. [4]–[7]). Therefore, such models have become the primary technique to understand the spread of the COVID-19.

The primary issue with such models is that they do not provide explainability in terms of the factors considered. This means that the reasoning behind such models must be concrete and a proper relationship between such factors or variables and the number of infected people must be identified clearly. This is because authorities need to know the direct relationship between factors and the disease spread. This paper aims to identify such relationships by using an Artificial Neural Network (ANN). The variables that present a cohesive relationship with the spread of the virus can then be selected as primary variables that go into more sophisticated models such as the compartmental models.

Therefore the primary contributions of this study can be understood as follows:

- Develop a dataset with meaningful variables and understand their applicability
- Apply an ANN to understand the efficiency in detecting the possible regions at high risk due to the coronavirus at the county level in the US
- Gain insights from the approach discussed and results obtained

The paper is organized as follows: Section II introduces the data sources and the design considerations for the ANN models; Section III explains the complete methodology of the approach; Section IV discusses the results obtained; and the conclusion is drawn in Section V.

II. Preliminaries

A. Data Collection

The primary dataset [8] used in this research has been obtained from the John Hopkins University (JHU) Dashboard for COVID-19 developed and maintained by JHU and the Esri Living Atlas Team. The main advantage of this dataset is that it has been aggregated at the county level in the US. The COVID-19 Community Mobility Reports [9] is a dataset developed by Google using anonymous location data. These reports present mobility statistics over time in several categories: retail & recreation; groceries; pharmacies; parks; transit stations; workplaces; and residential areas. This dataset is used to enhance the existing dataset by adding average mobility statistics for the time period considered. The variables in the combined dataset are presented and categorized in Table I.

TABLE I
Dataset Variables

Category	**Variables**
Demographics and Geography	Total Population Population by Age (13 categories) Population by Race (15 categories) Area of County Length of County
Economic Factors	Unemployment Rate Total Unemployed Median Household Income Median Household Income percent of State total Poverty Rate Poverty as percent of State
Medical Facilities	Number of Ventilators Number of Licensed Beds Number of ICU Beds Number of Staffed Beds
COVID Statistics	New Cases New Deaths Fatality Rate County Cases per 100,000 County Deaths per 100,000 State Fatality Rate State Confirmed State Deaths State Recovered State Tests
Mobility Factors	People in Grocery and Pharmacy Stores People in Parks People in Residential Areas

A primary issue with the combined dataset is that the data is not consistent among various counties. This means that some counties in the US have not collected the numbers as effectively as the other counties and furthermore, some counties have no data. This missing data can affect the performance of the ANN algorithm and therefore the data points for such counties have been eliminated to make the dataset more consistent and less sparse. Furthermore, since there are some counties with probably incorrect data due to inefficient data collection, it is assumed that they still do present the accurate trends in the rise and drop in the number of infected people.

B. Design Considerations for the ANN

To optimize the architecture of the network different activation functions, number of hidden layers, and loss functions were applied in order to achieve the best results. All variables were normalized to keep the loss function within the bounds of the computer memory. The ANN architecture is summarized in Table II.

1) Activation Function: Activation functions are primary components of a neural network and greatly affect the performance of a neural network. They determine whether a particular neuron should be activated or not, based on the inputs. A primary requirement is that activation functions must be computationally efficient as they are calculated across thousands or even millions of neurons for each data sample. The various available techniques are discussed below in order to select a desirable approach.

- The **Sigmoid / Logistic function** [10] has vanishing gradient which means for very high and very low values of X, there is almost no change to the prediction. This causes a vanishing gradient which leads to a network refusing to learn further, or being too slow to reach an accurate prediction. It is also computationally expensive.
- The **Hyberbolic tangent (tanh) function** [11] has the same drawbacks as sigmoid function. (i.e. the vanishing gradient problem) However, there is a key difference between TanH and Sigmoid. It is zero centered which makes it easier to model inputs that have strongly negative, neutral, and strongly positive values.
- The **Rectified Linear Unit (ReLU)** [12] is a computationally efficient technique and allows the network to converge faster. Although ReLu has a higly non-linear nature, it allows for backpropagation in an ANN. The main disadvantage with this function is that when an input approaches non-positive values, the gradient of the function becomes zero and therefore the network cannot perform back-propagation (Dying ReLU problem) which is highly undesirable.
- The **Leaky ReLU** [13] is an improvement over ReLU as it addresses the dying ReLU problem. It is achieved by having a small positive slope in the negative region which enables back-propagation. Despite this, it still has an issue that it doesn't provide consistent predictions for negative input values.

It is evident from the above discussion that Leaky ReLU is the best option to consider with one significant issue. Further modifications have been suggested to improve the consistency of the technique such as Parametric ReLU, Softmax, Swish-type, etc. However, for the purpose of this study, the mentioned issue is not significant as none of the variables (input/output) contain negative values. This makes Leaky ReLU the optimal choice and is therefore selected as the Activation Function.

2) Optimizer: The backbone of back-propagation algorithm is the optimization of weights and bias. Optimization has a very broad scope and a variety of techniques can be applied to develop an efficient system architecture. Common techniques that have been used over time for similar problems are the Gradient Descent, Stochastic Gradient Descent (SGD), and Adam's Method. These are explained below.

- **Gradient Descent (GD):** The GD method [14] is the rate of loss function with respect to the weights of the ANN model. The loss function can be a function of the mean square of the losses accumulated over the entire training dataset. Hence, the weights are updated at the end of each epoch. This results in avoidance of local minimum but requires a large computation time (or more number of epochs).
- **Stochastic Gradient Descent (SGD):** SGD method [14], on the other hand, is an improvement over the GD method. In this approach, the weights are updated after each training sample has been parsed.
- **Adam's Method:** Adam's method is a stochastic optimization technique that offers an adaptive learning rate for each parameter [15]. Parameters that would ordinarily be updated less frequently receive more regular updates with this technique. This increases the speed of the learning process.

A major benefit to Adam's method approach is that manual tuning becomes less important as the learning rate is adjusted automatically for all parameters. In comparison, SGD requires careful tuning (and possibly online adjustment) of learning rates. It is still necessary to select hyperparameters in the Adam's method but the performance is less sensitive to them as compared to the SGD method. Due to these advantages, Adam's method has been selected as the primary approach.

3) Validation Criteria: The criteria used to validate the obtained results is the *Smooth L1 Loss* technique, also known as the *Huber Loss* technique [16]. It is defined as follows:

$$\text{loss}(x,y) = \begin{cases} 0.5(x-y)^2, & \text{if } |x-y| < 1 \\ |x-y| - 0.5, & \text{otherwise} \end{cases} \tag{1}$$

In traditional techniques like mean square error (MSE) loss, we square the difference which results in a number that is much larger than the original number. This is a good approach when larger errors need to be highlighted. But it has a disadvantage that these high values result in exponentially increasing gradients.

Huber loss function utilizes a squared term only if the absolute error falls below 1. This way it is less sensitive to outliers than the MSE and in some cases prevents large changes in gradients. Therefore this technique avoids such situations and performs better when features have large values similar to the case presented in this paper.

4) Performance Metrics: R-squared is a statistical measure of how close the data are to the fitted regression line. It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression.

$$\mathbf{R}^2 = \frac{\text{Explained Variation}}{\text{Total Variation}} \tag{2}$$

This is one of the most popular techniques to measure the performance of ANNs and is highly suitable to regression-type problems.

TABLE II
ANN ARCHITECTURE

Parameter	**Value/Method**
Optimizer	Adam Optimizer
Learning Rate	0.001
Training Epochs	18,000
Hidden Layer	Leaky ReLU
Validation Criteria	Huber Loss Function
Performance Metric	R-squared

III. METHODOLOGY

The methodology is divided into 3 parts as follows:

- Model 1 design using complete set of variables.
- Reduction of variables using residual analysis.
- Model 2 design using reduced set of variables.

A. ANN: Model 1

Model 1 is designed to accommodate all the variables (55) discussed in Table I. The number of neurons in each layer is given in Table III. The number of neurons have been decided using manual verification using trial method. The parameters for the ANN model are given in Table II.

TABLE III
LAYER STRUCTURE: MODEL 1

Layer	**Number of Neurons**
Input	55
Hidden - 1	144
Hidden - 2	44
Hidden - 3	20
Output	1

B. Residual Analysis

The difference between the observed value of the dependent variable and the predicted value is called the residual. Each data point has one residual.

$$\theta = x - y \tag{3}$$

Residuals exhibit one important property that both the sum and the mean of the residuals are equal to zero (i.e. $\sum \theta = 0, \text{and } \bar{e} = 0$).

A residual plot is a graph that shows the relationship between the residuals and the independent variables. If the data points are randomly dispersed around the horizontal axis, a linear regression model is more appropriate for the data. On the other hand, a non-linear model is more applicable when the data is not randomly dispersed.

This allows us to select variables that are suited more for this kind of a problem and observe the variations in performance. Using a reduced dataset allows us to downsize the required dataset which has a number of advantages discussed in Section III-C.

In order to define the model 2, the reduced dataset needs to be defined. Selection of a variable is dependent on the residual plot of that variable as explained below.

- A weak variation between variable and output (bad variable) (See Figure 1)
- A strong variation between variable and output (good variable) (See Figure 2)

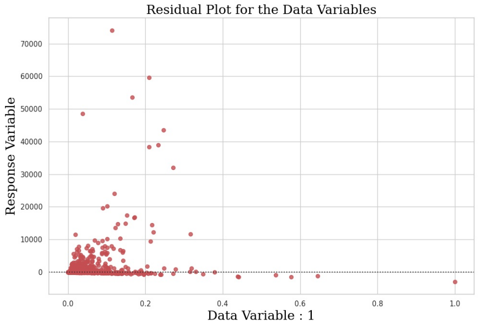

Fig. 1. Bad Residual Plot

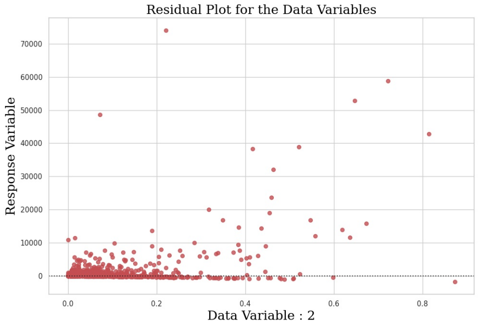

Fig. 2. Good Residual Plot

This criteria is defined to separate the good variables and bad variables. Now, the good variables are used to define the dataset for Model 2.

The new, reduced set of variables obtained from residual analysis includes - Deaths per 100,000 population, Overall Poverty, State Fatality Ratio, Median Household Income, State Confirmed Cases, State Confirmed Deaths, State Number of Tests, and People in Grocery and Pharmacy Stores.

C. ANN: Model 2

The number of neurons in each layer is given in Table IV. Similar to Model 1, the number of neurons in Model 2 have been decided using manual verification using trial method. The parameters for the ANN model are given in Table II.

This model offers a number of advantages compared to the model that uses the complete dataset.

- Allows for a select number of variables to be used for other models by reducing the redundant variables
- Highlights the importance of data collection in a specific and focused manner

TABLE IV
LAYER STRUCTURE: MODEL 2

Layer	**Number of Neurons**
Input	8
Hidden - 1	20
Hidden - 2	7
Hidden - 3	5
Output	1

IV. RESULTS

Figures 3 and 4 show the regression plots during training of Models 1 and 2. Here, it is evident that the performance of the model using whole data set is better than with the reduced data set. This is expected as the variable selection criteria is not explicitly defined.

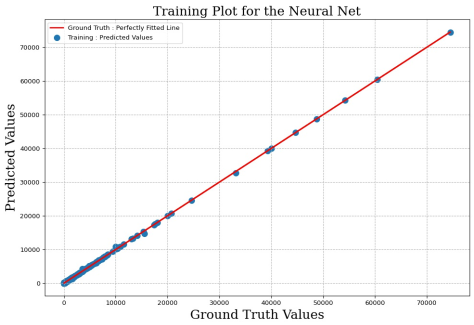

Fig. 3. Regression Plot - Training of Model 1

Figures 5 and 6 depict the regression plot after training and we can see trends similar to the training plots.

The results highlight the fact that downsizing the dataset can lead to loss of performance when an explicit criteria is undefined. Nonetheless, it can also be observed that the reduced number of variables do show a strong correlation with the output variable.

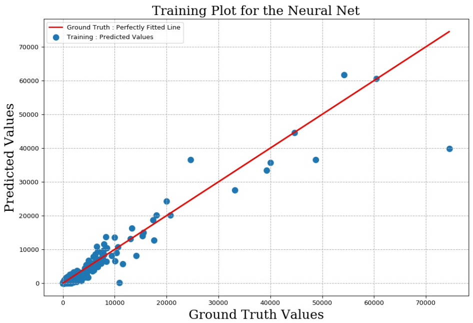

Fig. 4. Regression Plot - Training of Model 2

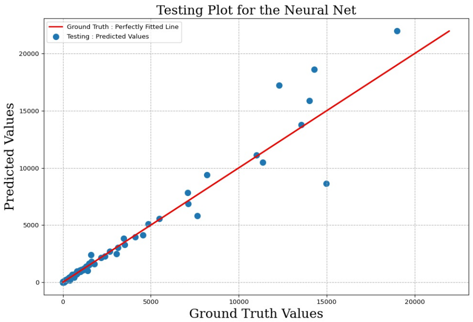

Fig. 5. Regression Plot - Testing of Model 1

V. Conclusion

The results show that the reduced dataset of variables can be generated using the suggested method. However, an explicit criteria can be defined to automatically define the reduced dataset from any complete dataset. This also presents application in other application areas in data science where variable reduction is needed during the data pre-processing without any huge computational cost. This provides an interesting direction for future study. Furthermore, the method proposed is advantageous as a specific set of variables can be defined for the development of infectious disease models. In addition, it helps identify important human factors which can be monitored regularly to identify the spread of an infectious disease such as COVID-19 in the future. This can prove instrumental in curbing the spread and working on reducing the effect of those human factors on the spread of the disease. This, in turn, can help optimize the process of lockdown measures, and reopening ~~and~~ timelines.

Acknowledgment

The authors would like to thank the NASA International Space Apps Challenge 2020 and everyone involved in this event for the call to action that motivated our research. The authors would also like to acknowledge the efforts of Dr. Gautam Pillay and Mr. Scott Peterson who have guided us constantly in the right direction.

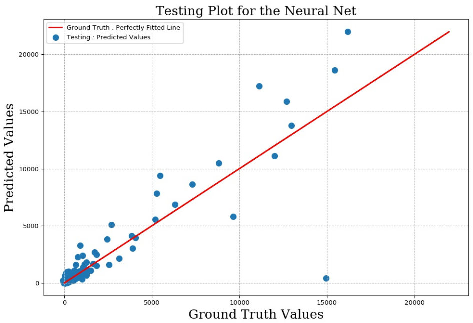

Fig. 6. Regression Plot - Testing of Model 2

References

[1] G. Bonaccorsi, F. Pierri, M. Cinelli, A. Flori, A. Galeazzi, F. Porcelli, A. L. Schmidt, C. M. Valensise, A. Scala, W. Quattrociocchi, and F. Pammolli, "Economic and social consequences of human mobility restrictions under COVID-19," *Proceedings of the National Academy of Sciences*, vol. 117, no. 27, pp. 15 530–15 535, 2020. [Online]. Available: https://www.pnas.org/content/117/27/15530

[2] W. O. Kermack, A. G. McKendrick, and G. T. Walker, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A, Containing Papers of a Mathematical and Physical Character*, vol. 115, no. 772, pp. 700–721, 1927. [Online]. Available: https://royalsocietypublishing.org/doi/abs/10.1098/rspa.1927.0118

[3] J. L. Aron and I. B. Schwartz, "Seasonality and period-doubling bifurcations in an epidemic model," *Journal of Theoretical Biology*, vol. 110, no. 4, pp. 665–679, 1984. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0022519384801502

[4] B. R. E. Rowthorn and F. Toxvaerd, "The Optimal Control of Infectious Diseases Via Prevention and Treatment," *CEPR Discussion Paper No. DP8925*, 2012. [Online]. Available: https://ssrn.com/abstract=2034143

[5] M. S. Eichenbaum, S. Rebelo, and M. Trabandt, "The macroeconomics of epidemics," National Bureau of Economic Research, Working Paper 26882, March 2020. [Online]. Available: http://www.nber.org/papers/w26882

[6] C. Gollier, "Cost-benefit analysis of age-specific deconfinement strategies," *Covid Economics*, vol. 24, pp. 1–31, 2020.

[7] C. A. Favero, A. Ichino, and A. Rustichini, "Restarting the Economy While Saving Lives Under COVID-19," 2020. [Online]. Available: https://ssrn.com/abstract=3580626

[8] Centers for Civic Impact, JHU, "US Counties Cases," 2020. [Online]. Available: https://services9.arcgis.com/6Hv9AANartyT7fJW/arcgis/rest/services/USCounties_cases_V1/FeatureServer/0

[9] Google, "Community Mobility Reports," 2020. [Online]. Available: https://www.google.com/covid19/mobility/

[10] J. Han and C. Moraga, "The influence of the sigmoid function parameters on the speed of backpropagation learning," in *From Natural to Artificial Neural Computation*, J. Mira and F. Sandoval, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1995, pp. 195–201.

[11] Y. A. LeCun, L. Bottou, G. B. Orr, and K.-R. Müller, *Efficient BackProp*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 9–48. [Online]. Available: https://doi.org/10.1007/978-3-642-35289-8_3

[12] R. H. R. Hahnloser, R. Sarpeshkar, M. A. Mahowald, R. J. Douglas, and H. S. Seung, "Digital selection and analogue amplification coexist in a cortex-inspired silicon circuit," *Nature*, vol. 405, no. 6789, pp. 947–951, Jun 2000. [Online]. Available: https://doi.org/10.1038/35016072

[13] A. L. Maas, "Rectifier nonlinearities improve neural network acoustic models," 2013.

[14] S. Ruder, "An overview of gradient descent optimization algorithms," 2016.

[15] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," 2014.

[16] P. J. Huber, "Robust estimation of a location parameter," *Ann. Math. Statist.*, vol. 35, no. 1, pp. 73–101, 03 1964. [Online]. Available: https://doi.org/10.1214/aoms/1177703732