

Combining machine learning and non-linear dynamics modeling to understand COVID-19 risk factors

Djordjevic M.^{1*}, Salom I.², Djordjevic M.², Sofija M.¹, Rodic A.¹, Milicevic M.¹, Tumbas M.¹, Zigic D.²

¹ Quantitative Biology Group, Faculty of Biology, University of Belgrade, Belgrade, Serbia

² Institute of Physics Belgrade, University of Belgrade, Belgrade, Serbia

* dmarko@bio.bg.ac.rs

Key words: infectious diseases, dynamical epidemics models, COVID-19 transmissibility, COVID-19 severity, feature selection, Random Forest, Gradient Boost

Motivation and Aim: Many studies have been investigating the relationship between COVID-19 transmissibility and severity in population and different sociodemographic, health-related, and environmental factors. Confirmed case counts (infected and fatalities) were commonly used as a measure of disease transmissibility and severity. However, a major limitation of these studies is that they do not adequately account for complex disease dynamics, influenced by, e.g., significant differences in control measures, testing policies, and points on the epidemic curve associated with different geographic regions. Another difficulty is appropriately controlling the effects of multiple potentially important factors, due to both their mutual correlations, limited dataset, and potential non-linearities and interactions between these factors. To address these difficulties, we use epidemiological modeling to i) Estimate the basic reproduction number (R_0) as a robust measure of the disease severity directly from observed infection count data [1], ii) Propose a disease severity measure independent of the disease transmission dynamics (i.e., the basic reproduction number) that has a direct mechanistic interpretation. The derived measure corresponds to the ratio of population-averaged mortality and recovery rates (m/r) [2]. We use these two measures to assess demographic, medical, meteorological, and environmental factors associated with the disease severity. For this, we use machine learning regressions that employ different feature selection methods [1–3]. To reduce dimensionality without complicating the variable interpretation and partially decorrelate the variable set, we employ Principal Component Analysis on subsets of mutually related (and highly correlated) predictors. Finally, we investigate how the results change when the data are analyzed on a smaller special resolution (specifically, when going from USA states to USA counties).

Results: We analyze COVID-19 transmissibility [1] and severity [2, 3] on case count data from US states, where we assemble many potentially relevant variables. We use both linear regressions with regularization and feature selection (Lasso and Elastic Net) and non-parametric methods based on ensembles of weak-learners (Random Forest and Gradient Boost). Through these substantially different approaches, we robustly obtain that $PM_{2.5}$ is a major predictor of R_0 in USA states, while factors such as other pollutants, prosperity measures, population density, chronic disease levels, and possibly racial composition are also significant [1]. For the disease severity, without using prior knowledge from clinical studies, we recover significant predictors known to be risk factors, i.e., age, chronic diseases, and racial factors. Additionally, we identify long-term pollution exposure and population density as not widely recognized (though for the

pollution previously hypothesized) predictors of the disease severity (see Fig. 1 below) [2]. Analysis at the county level also identifies race (proportion of the Black population in particular) as a major factor strongly associated with the disease severity [3].

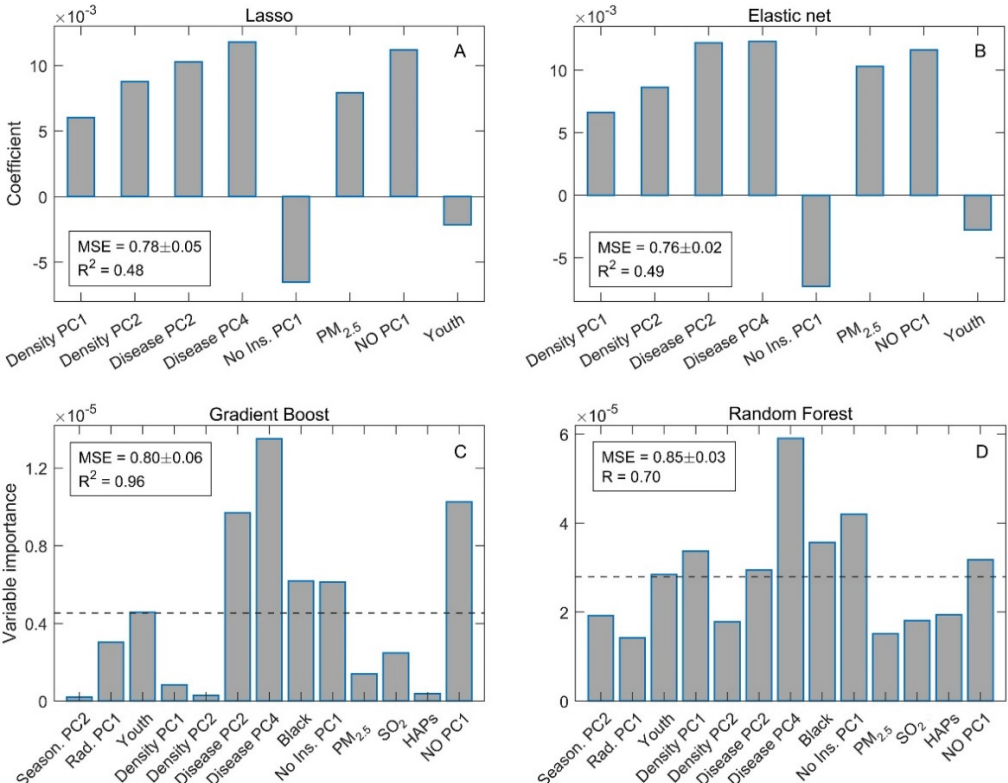


Fig. 1. Machine learning analysis of COVID-19 severity risk factors. A) Lasso and B) Elastic Net, C) Gradient Boosting, and D) Random Forest regressions. For A) and B), the height of bars corresponds to the value of the regression coefficients. For C) and D), the height of bars corresponds to estimated predictor importance. Variable names are indicated on the horizontal axes. Figure adapted from Ref. [2]

Conclusion: The measures of the disease progression inferred or derived from the epidemics dynamical modeling proved useful for inferring risk determinants of COVID-19. When applied to data through machine learning approaches, they lead to robust and plausible results that aid understanding COVID-19 risks. Overall, we find that a combination of non-linear dynamics and machine learning is advantageous in analyzing the disease progression, which may also be applied to potential future disease outbursts.

Acknowledgements: This study is supported by the Ministry of Education, Science and Technological Development of the Republic of Serbia.

References

1. Djordjevic M., Rodic A., Salom I., Zigic D., Milicevic O., Ilic B., Djordjevic M. A systems biology approach to COVID-19 progression in a population. *Adv Protein Chem Struct Biol.* 2021;127:291-314.
2. Djordjevic M., Djordjevic M., Ilic B., Stojku S., Salom I. Understanding Infection Progression under Strong Control Measures through Universal COVID-19 Growth Signatures. *Global Challenges.* 2021;5(5):2000101. doi: 10.1002/gch2.20200010.
3. Tumbas M., Markovic M., Salom I., Djordjevic M. A large scale machine learning study of demographic factors contributing to COVID-19 severity, submitted. 2021.