Regional characteristics of the second wave of SARS-CoV-2 infections and COVID-19 deaths in Germany: a machine learning approach

Gabriele Doblhammer, Daniel Kreft, Constantin Reinke

The second wave of SARS-CoV-2 infections began in Germany in October 2020, increased exponentially in November, and remained at high levels well into December, despite various regulatory measures beginning in September 2020 and a lockdown beginning in early November 2020.

The aim of this study was to identify the key features explaining SARS-CoV-2 infections and COVID-19 deaths during the upswing of the second wave in Germany. There is a consensus that infections and deaths have affected lower social groups the hardest (for a review see [1], mainly due to their higher mobility during the pandemic and their lower capacity for social distancing [2].

U.S. and U.K. studies were the first and most prominent to indicate that persons of ethnic minorities were at increased risk for COVID-19 compared with whites (for a systematic review see [3]. Comorbidities, barriers related to language, health seeking and health care, cramped housing, risky work and working conditions [4, 5] have been identified as risk factors and vulnerabilities leading to higher virus exposure. In Germany, migrants are highly represented in occupations with system relevance and thus a higher potential exposition to the virus, such as cleaning workers, workers in food production, or nursing of elderly.

While social distancing is essential to contain the spread of COVID-19, not everyone is willing to comply with social distancing measures. From the U.S., there are reports based on debit card transaction data that Democrats were more likely to switch to remote spending after government orders were implemented [6]. Also for the U.S., political conservatism inversely predicted compliance with behaviors aimed at preventing the spread of COVID-19 [7].

Data

We downloaded data (January 26, 2021) from the Robert Koch Institute [8] which provides information on COVID-19 diagnoses and deaths by sex, age (age groups: 0-4, 5-14, 15-34, 35-59, 60-79, 80+), and 401 counties (NUTS3 region). Population size on the county level was derived from the DESTATIS regional database at the end of the year. For characteristics of the counties we used the INKAR (Indikatoren und Karten zur Raum- und Stadtentwicklung) database (2020), the 2011 German census, emission data from the German Environment

Agency Database (UBA), main diagnoses in hospitals by place of residence in 2017 from the Regional Database of the Statistical Offices of the Federation and the Länder, and the international COVID-19 incidence rates from the European Center for Disease Prevention and Control. Latitude and longitude were defined in terms of the centers of the county capitals.

Analysis strategy and Method

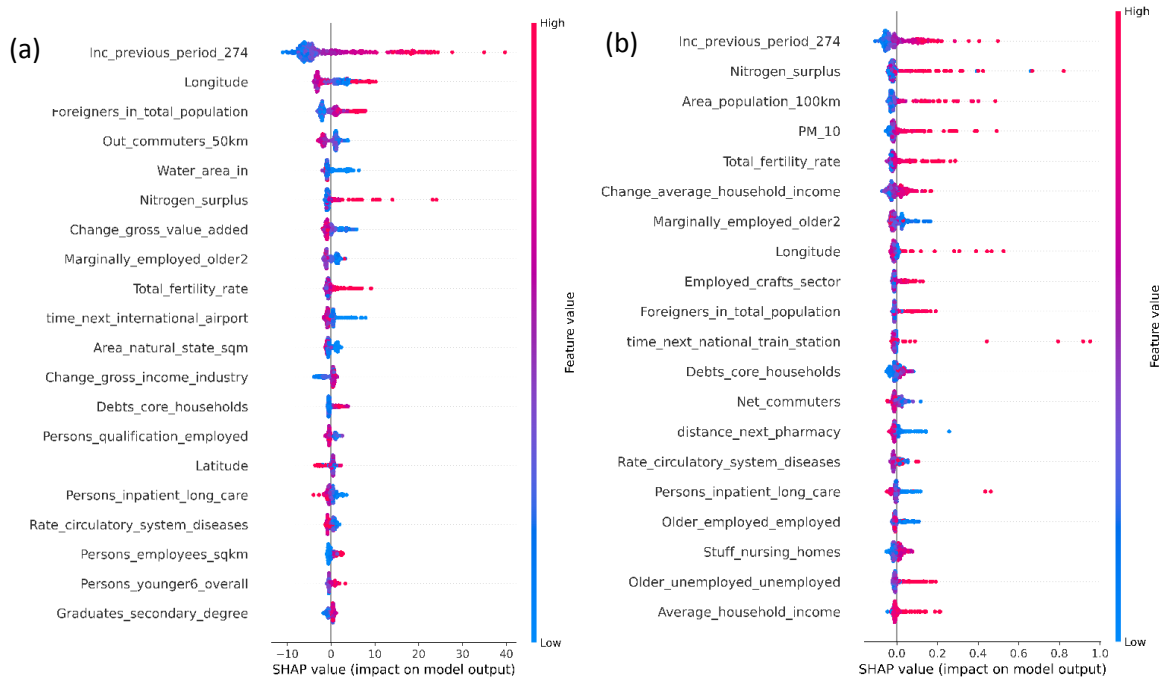
Our analysis strategy consisted of three steps: First, we trained gradient boosting models to predict the age-standardized incidence and death rates for each period with the 155 characteristics of the counties; these characteristics are termed. Gradient boosting models were trained using the CatBoostRegressor from the CatBoost algorithm. As an alternative, we used the random forest regressor from the Scikit-learn module in Python with 5,000 trees. We kept all other hyperparameters at their default values. We calculated the R^2 and root mean squared (RMSE) errors to evaluate how well the models fit the data. Second, we used Shap values to explore the importance of the features and third, we characterized the 20 most prominent features in terms of negative/positive associations with each of the two outcome variables. We categorized the top 20 associations identified by the Shap values into twelve categories depicting the correlation between the feature and the outcome.

Results

The age-standardized incidence and death rates changed over time, as did the key features. At the beginning of the second wave (Oct. 01-15), the overall low incidence was comparatively elevated in both high and low SES regions (Figure 1 a,b) as indicated by the Shap values; with the exponential increase in infections, low SES regions were more heavily affected; no single feature was associated with high incidence in high SES (8 of 20 features related to low SES); Nov. 01-15 (5 of 20 features related to low SES). During the peak period (Nov: 16-30; Dec: 01-15), infections again spilled over from low to high SES regions. COVID-19 deaths were correlated with both low and high SES regions in all periods, with generally higher correlation with low SES regions (Figure 2b).

Taking the first twenty features according to their Shap values, we grouped them into the twelve categories outlined above. We counted the number of features in each category and found that features related to SES, urbanity/density, and health were present in all time periods; those representing the connectedness of a region were present in the period from mid-October to mid-November and again in December.

Figure 1: Boosting Model: SHAP summary plots of the first twenty features (a) age-standardized incidence, (b) age-standardized death rates



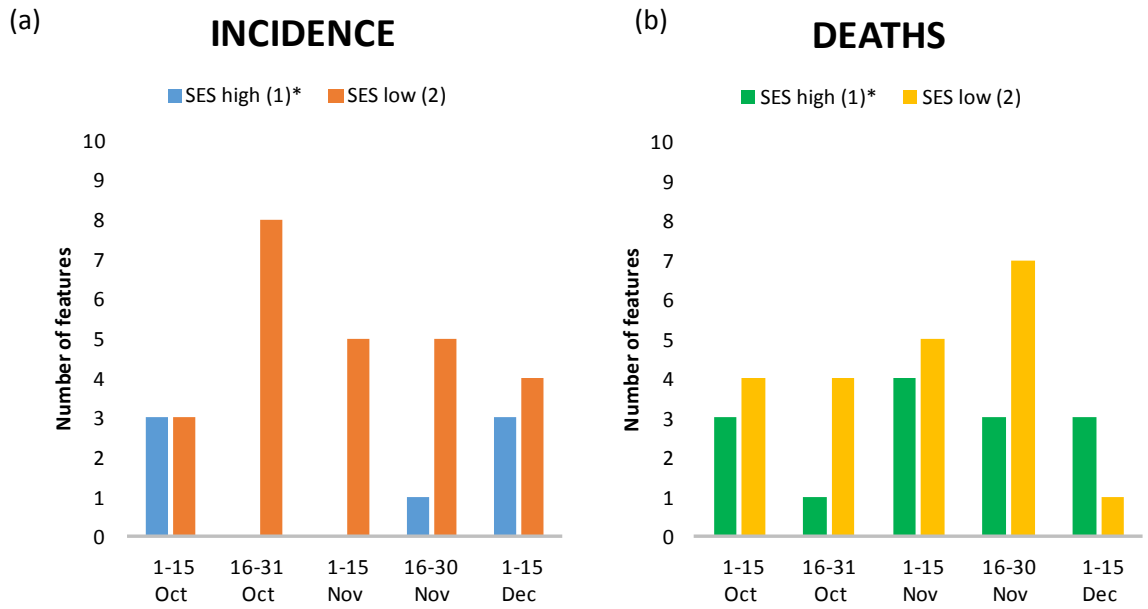
The summary plot combines feature importance with feature effects. Each point on the summary plot is a Shapley value for a feature and a county. The position on the y-axis is determined by the feature and on the x-axis by the Shapley value. The color represents the value of the feature from low to high. Overlapping points are jittered in y-axis direction, to get a sense of the distribution of the Shapley values per feature. The features are ordered according to their importance. E.g. Low values of the age-standardized incidence in the previous period (inc_previous_period_274) are correlated with low values in the age-standardized incidence of the current period (a). High values of longitude (Longitude) are correlated with low values of the age-standardized incidence of the current period (a). High levels of "Nitrogen surplus per agricultural area in kg/ha in 2016" (Nitrogen_surplus) are correlated with high age-standardized death rates in the current period (b).

Features related to need for care started to show up in the second half of November and in December, while those related to migration were present in October and the first half of November. Features reflecting values and norms in terms of regional voting behavior were present in all periods, as did those characterizing the (age) structure and aging process in a region.

Discussion

Our study showed that an ecological approach using explainable machine learning methods can help shed more light on the regional infection patterns of COVID-19 in Germany. Ecological analyses have their place in stimulating innovation in a rapidly evolving field of research where individual data are not yet available. Although ecological analysis cannot provide insight into mechanisms and does not allow for inference regarding individuals, it can highlight potential drivers.

Figure 2: Number of features in the top 20 showing the relationship between low and high SES, and incidence (a) and death rates (b), by time period.



A number of regional characteristics were crucial for the increase in infections in the second wave. As in the first wave, they moved from high to low SES regions. Risky working conditions with reduced opportunities for social distancing, a high burden of chronic disease, and residence in nursing homes may underlie this concentration in low-SES regions. In addition, regional patterns of voting behavior were associated with infections and deaths, possibly indicating norms and values associated with nonadherence to Corona measures. To further elucidate these findings, we urgently need more individual-level data.

References

1. Wachtler, B.; Michalski, N.; Nowossadeck, E.; Diercke, M.; Wahrendorf, M.; Santos-Hövener, C.; Lampert, T.; Hoebel, J. Socioeconomic inequalities and COVID-19 – A review of the current international literature. *Journal of Health Monitoring* **2020**, *5*, 3–17, doi:10.25646/7059.
2. Weill, J.A.; Stigler, M.; Deschenes, O.; Springborn, M.R. Social distancing responses to COVID-19 emergency declarations strongly differentiated by income. *Proceedings of the National Academy of Sciences* **2020**, *117*, 19658–19660, doi:10.1073/pnas.2009412117.
3. Sze, S.; Pan, D.; Nevill, C.R.; Gray, L.J.; Martin, C.A.; Nazareth, J.; Minhas, J.S.; Divall, P.; Khunti, K.; Abrams, K.R. Ethnicity and clinical outcomes in COVID-19: a systematic Review and Meta-analysis. EClinicalMedicine 2020, doi:10.1016/j.eclinm.2020.100630
4. Reid, A.; Ronda-Perez, E.; Schenker, M.B. Migrant workers, essential work, and COVID-19. American Journal of Industrial Medicine 2021, 64, 73–77.
5. Middleton, J.; Reintjes, R.; Lopes, H. Meat plants - a new front line in the covid-19 pandemic. *BMJ* **2020**, *370*, m2716, doi:10.1136/bmj.m2716.
6. Painter, M.; Qiu, T. Political beliefs affect compliance with covid-19 social distancing orders. *Journal of Economic Behavior and Organization, Forthcoming* **2020**, doi:10.2139/ssrn.3569098.
7. Müller, S.; Rau, H.A. Economic preferences and compliance in the social stress test of the COVID-19 crisis. *Journal of Public Economics* **2021**, *194*, 104322, doi:10.1016/j.jpubeco.2020.104322.
8. Robert Koch Institute; ESRI. RKI Corona Landkreise. Available online: https://npgeo-corona-npgeo-de.hub.arcgis.com/datasets/917fc37a709542548cc3be077a786c17_0?selectedAttribute=cases_per_population (accessed on 26 January 2021).