**Modern Data Analytics:   
Final Project Report - Group Sweden**

Prof. Jan De Spiegeleer

7 June 2022



| **Name** | **Student number** | **Email address** |
| --- | --- | --- |
| Kinga Nowak | r0862670 | kingaanna.nowak@student.kuleuven.be |
| Emma Kessenich | r0866746 | emmasofie.kessenich@student.kuleuven.be |
| Fan Huang | r0822280 | fan.huang@student.kuleuven.be |
| Sara Rutten | r0750751 | sara.rutten@student.kuleuven.be |
| Ilias Willems | r0747852 | ilias.willems@student.kuleuven.be |
| Peter Day | r0866276 | peter.day@student.kuleuven.be |

Introduction and problem statement  
After Covid-19 was first discovered in Wuhan, it spread like wildfire, quickly transforming into a pandemic. Our analysis focuses on the evolution of this disease in the United States by trying to answer the research question: “How did Covid-19 evolve in the US between 2020-01-21 and 2022-04-20?”. To narrow down this general question, we formulated several sub-questions: “Which clusters of states/counties in the US evolved similarly and which features influence these clusters?”, “How can we predict the evolution of Covid-19 cases in the US?” and “Which counties in the US can be classified as hotspots and which features explain this classification?”.

Pipeline  
Before starting our analysis, data was retrieved from different sources. The initial dataset of the NY Times, containing the Covid-19 cases and deaths at the county level, was chosen as the central point of our analysis (NY Times, 2022). To investigate which variables had an influence on the Covid-19 evolution, other data was retrieved related to vaccination, commuting flows and transportation, Covid policy measures, political opinion, waste water, climate, health related variables and variables related to demographic characteristics such as age, population, education, income and unemployment.  
 In a second phase, the data preprocessing was done. Since each dataset is different, no general pipeline could be applied but approaches specifically tailored to a dataset had to be used. First, for our central dataset on daily Covid-19 cases and deaths, we calculated the number of cases per week and computed the infection and death rates per 100,000 inhabitants. Next, to generate features for our analyses, the Covid-19 policy measures data was cleaned such that only dates that mark changes in policies concerning mask use, schools closures and vaccination were kept. Additionally, the time-dependent vaccination data was preprocessed as well by replacing missing values with the last recorded value for a particular county and the absolute numbers of administered vaccine doses were transformed into vaccination rates per county. To measure political opinion, the percentages of votes cast for Democratic and Republican candidates were calculated, and for all features concerning health, transportation and demographics, missing values for single counties were imputed using the corresponding state-wide mean. As the waste water data is reported as a weekly rolling average but the number of cases are measured daily, the weekly data is turned into daily data, by giving the three days before and after the reporting day the same value as the reported value.   
 The next preprocessing step concerns the dataset on commuting flows between counties. After removing missing values, commuting flows were analyzed in Neo4j, both on state and county level. In both cases, Neo4j was used to construct a directed graph where every node represents a state/ county and two nodes are connected by an edge that is weighted by the number of commuters. The degree centrality (at state-level) as well as closeness centrality, betweenness centrality and pagerank score (at county-level) were calculated as measures of states’/ counties’ interconnectedness.   
 To use for our analyses, all non-time-dependent features we constructed were merged by state and county. Moreover, we defined five periods across which we compared the evolution of Covid-19 in the US (Jan-June 2020, June-Dec 2020, Dec-June 2021, June-Dec 2021 and Dec-Apr 2022) and we aggregated all time-dependent features per period.   
 After preprocessing the data, our analyses could be performed, structured according to the three questions we defined above. The first part is related to finding communities and clusters. This was done based on clustering algorithms such as K-means and spectral clustering in a pipeline (at county-level, based on weekly infection rates), as well as using Label propagation in Neo4j (at state-level, based on daily reported seven-day trailing averages). For the latter, we constructed a graph based on correlations between the reported averages of the number of cases/deaths and created an undirected connection between two states if their correlation is found to be larger than 0.7. Second, to predict the future evolution of Covid-19 cases, we zoom in on two time-dependent variables: we performed a linear model to predict case evolution based on the virus concentration in the waste water and an SVM model was fitted to predict the infection rates based on the Covid-19 policy measures. Lastly, to understand more comprehensively which features explain high infection and death rates, a random forest model was fitted to determine, using Shapley plots, which features had an influence on the classification of a county as a hotspot (above average number of cases/ deaths) or not. For more information on these analyses, we refer to the app <https://app-modern-data-analytics.herokuapp.com/>.

Results  
Concerning the first question, the detection of clusters of similar Covid-19 evolution, we find that although the results differ from period to period, some interesting remarks can be made. Based on community detection in Neo4j, vaccinations seem to have an effect on the clustering solution of the Covid-19 cases in period 4. Besides, the clustering solution for the cases in this period turned out to be very similar to the clusters that were found with K-means clustering (after hyperparameter optimization). Also the clustering solution related to Covid-19 deaths differed from period to period, often concluding that the proportion of people that are uninsured influences community membership.  
 With regard to our second question, the prediction of future Covid-19 evolution, we found that it is difficult in general to predict the evolution of Covid-19 due to the large number of influencing factors, but virus concentration in the waste water turns out to be a useful predictor. Using a lag of 11 days, there is a very strong association between the concentration of Covid-19 in waste water and the number of Covid-19 cases. Yet, it should be noted that the waste water concentration is an early symptom, rather than a true explanation of the evolution of Covid-19.   
 To investigate the explanatory factors behind Covid-19 hotspots for our third question, it turned out that the importance of the factors changed over the different periods, except for measures of social disadvantages that consistently play key roles. Related to vaccination, an interesting remark can be made. Starting when vaccinations were rolled out to the public (period 3), vaccinations are always among the top influencing factors on both cases and deaths hotspots. However, looking at the cases in periods 4 and 5, a higher vaccination rate is associated with a higher probability of being a hotspot. Hence, although vaccination does seem to have an influence on the Covid-19 cases, which was also seen in the clustering solutions, it is not the case that a higher vaccination rate leads to a lower hotspot probability for cases (except in period 3). This could be due to changing behavior or relaxing other measures because of vaccination. However, interestingly, a higher vaccination rate is always associated with a lower probability of being a death hotspot (for periods 3, 4 & 5). This points into the direction of a lower probability of being severely ill if the vaccination rate is higher.   
 For more results, we refer to the app <https://app-modern-data-analytics.herokuapp.com/>.

Conclusion

Understanding and predicting the evolution of Covid-19 is not an easy task. However, by combining data from various sources, we were able to provide answers to our research questions. After clustering different states/counties based on their similarity in the evolution of Covid-19, the vaccination and uninsurance rates - amongst others - are seen to be important factors when it comes to modeling the cases and Covid-19 related deaths in the different clusters. Predicting the future evolution of the disease proves to be difficult, although the concentration of Covid-19 virus particles in wastewater can be used to make highly accurate short term predictions. In terms of underlying explanatory factors that affect the pandemic’s evolution, the influencing factors are manifold and vary over time as shown by our analyses to locate and characterize infection/ death hotspots, but in particular measures of social disadvantages and vaccination play key roles.

# References

Biobot Analytics. (2022, April 20). *Covid19-wastewater-data: wastewater by county [CSV file].* Retrieved from <https://github.com/biobotanalytics/covid19-wastewater-data/blob/master/wastewater_by_county.csv>.

Centers for disease control and prevention. (2022, March 31). *COVID-19 Vaccinations in the United States,County [CSV file]*. Retrieved from <https://data.cdc.gov/Vaccinations/COVID-19-Vaccinations-in-the-United-States-County/8xkx-amqh>.

Climate Change Knowledge Portal.(2020). *Mean temperature of the US + subnational units (monthly and annually) [CSV file]*. Retrieved from <https://climateknowledgeportal.worldbank.org/download-data>.

Harvard Dataverse. (2022, March 18). *County Presidential Election Returns 2000-2020 [tab file]*. Retrieved from https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/VOQCHQ

Health data. (2014). *USA County mortality risk 1980-2014 [xlsx file]*. Retrieved from <https://ghdx.healthdata.org/sites/default/files/record-attached-files/IHME_USA_COUNTY_LE_MORTALITY_RISK_1980_2014_NATIONAL_XLSX.zip> .

KFF. (2019). *Health Insurance Coverage of the Total Population [CSV file*]. Retrieved from <https://www.kff.org/other/state-indicator/total-population/?currentTimeframe=0&sortModel=%7B%22colId%22:%22Location%22,%22sort%22:%22asc%22%7D>.

Lilian Kilduff. (2021, December 22). *States Ranked by Percent of Population Age 65 or Older, 2020 [CSV file]*. Retrieved from <https://www.prb.org/resources/which-us-states-are-the-oldest/>.

NYTimes. (2022, April 11). *Covid-19-data [CSV file].* Retrieved from [GitHub - nytimes/covid-19-data: An ongoing repository of data on coronavirus cases and deaths in the U.S.](https://github.com/nytimes/covid-19-data)

NYTimes. (2020). *Mask-Wearing Survey Data [CSV file]*. Retrieved from <https://github.com/nytimes/covid-19-data/tree/master/mask-use>

OpenIntro. (2019). *United States Counties [CSV file]*. Retrieved from <https://www.openintro.org/data/?data=county_complete>

Oxford University. (2022, April 26). *Covid Policy Tracker [CSV file]*. Retrieved from <https://github.com/OxCGRT/covid-policy-tracker>.

United States Census Bureau. (2011). *Land Area [xls file]*. Retrieved from <https://www.census.gov/library/publications/2011/compendia/usa-counties-2011.html#LND>

United States Census Bureau. (2015). *Residence County to Workplace County Commuting Flows for the United States and Puerto Rico Sorted by Workplace Geography: 5 year ACS, 2011-2015 [xlsx file]*. Retrieved from <https://www.census.gov/data/tables/2015/demo/metro-micro/commuting-flows-2015.html>.

United States Census Bureau. (2020). *National Population Total 2010-2020: Annual Resident Population Estimates for States and Counties (CO-EST2020) [CSV file]*. Retrieved from https://www2.census.gov/programs-surveys/popest/datasets/2010-2020/counties/totals/

USDA. (2022). *State FIPS Codes [CSV file]*. Retrieved from <https://www.nrcs.usda.gov/wps/portal/nrcs/detail/?cid=nrcs143_013696>.

U.S. Department of Transportation Bureau of Transportation Statistics. (2021, August 24). *County Transportation Profiles [CSV file]*. Retrieved from <https://data.bts.gov/Research-and-Statistics/County-Transportation-Profiles/qdmf-cxm3>.

World Population Review. (2022). *Poverty Rate by State [CSV file]*. Retrieved from <https://worldpopulationreview.com/state-rankings/poverty-rate-by-state>.

World Population Review. (2022). *US States - Ranked by Population 2022 [CSV file]*. Retrieved from <https://worldpopulationreview.com/states>.

World Population Review. (2022). *List of State Abbreviations [CSV file]*. Retrieved from <https://worldpopulationreview.com/states/state-abbreviations>.

Appendix: Description of measures used

| **Variable** | **Description** | **Used in which analysis** |
| --- | --- | --- |
| **Covid cases & deaths** | | |
| Daily Covid-19 cases | The number of daily newly reported Covid-19 cases in the United States as a whole. | Simple linear model |
| Weekly Covid-19 infection rate | The infection number calculated per week and per state. (The infection number is calculated as the number of cases in the next week divided by the number of cases in the present week.) | SVM model |
| Seven-day trailing average (cases) | The seven-day rolling average of Covid-19 related cases for each day, aggregated by state. | Community detection (LPA) |
| Seven-day trailing average (deaths) | The seven-day rolling average of Covid-19 related deaths for each day, aggregated by state. | Community detection (LPA) |
| Weekly Covid-19 cases | The number of Covid-19 cases, aggregated by week, per county. | K-means model |
| Hotspot per period (cases) | Binary target per county and period whether a county’s average number of cases per 100k in a period was above the period’s average. | Random forest models |
| Hotspot per period (deaths) | Binary target per county and period whether a county’s average number of deaths per 100k in a period was above the period’s average. | Random forest models |
| **Vaccination & health-related features** | | |
| Complete rate per period | Average per county and period what percentage received a complete vaccine dose. Computed based on the average number of complete vaccine doses administered per county and period. | Random forest models  Community detection (LPA)  K-means model |
| Booster rate per period | Average per county and period what percentage received a booster vaccine dose. Computed based on the average number of complete vaccine doses administered per county and period. | Random forest models |
| Uninsured | Percentage of population that is uninsured in 2019, per county. | Random forest models  K-means model |
| Insured per state: uninsured | Percentage of population that is uninsured in 2019, per state | Community detection (LPA) |
| Life expectancy | Life expectancy in 2014, per county. | Random forest models  K-means model |
| **Commuting & transportation** | | |
| Degree centrality per state | Baseline metric of connectedness based on the commuting flow between states. | Community detection (LPA) |
| Closeness centrality per county | Centrality of a county based on the commuting flow between counties. | Random forest models |
| Betweenness centrality per county | Control points based on the commuting flow between counties. | Random forest models |
| Pagerank score per county | Overall influence of a county based on the commuting flow between counties. | Random forest models  K-means model |
| Airports | Number of primary and commercial airports per county. | Random forest models  K-means model |
| **Waste water** | | |
| Effective virus concentration rolling average per day | SARS-CoV-2 virus concentration present in samples of wastewater. Sampling locations with larger populations will contribute more to the value (population-weighted average). | Simple linear model |
| **Covid policy measures** | | |
| Mask use policy changes | Dates on which mask use policies changed. | SVM model |
| School closure policy changes | Dates on which school closure policies changed. | SVM model |
| Vaccination availability changes | Dates on which vaccination availability changed. | SVM model |
| Average mask use per period | Computed average per state and period how long policies concerning mask use were in place. | Random forest models |
| Average school closures per period | Computed average per state and period how long policies concerning school closures were in place. | Random forest models |
| Average vaccination per period | Computed average per state and period how long policies concerning vaccination were in place. | Random forest models |
| **Climate** | | |
| Mean temperature per state | Mean temperature per state in 2020. | Random forest models |
| Mean temperature per period | Computed mean temperature per period (Dec-June or June-Dec) in the US in 2020. | Random forest models |
| **Political opinion** | | |
| Vote Republican | Computed percentage of votes cast for a Republican candidate during 2020 presidential elections in a county. | Random forest models  K-means model |
| Vote Democrat | Computed percentage of votes cast for a Democratic candidate during 2020 presidential elections in a county. | Random forest models  K-means model |
| **Social demographic characteristics[[1]](#footnote-0)** | | |
| Senior population | Percentage of state population older than 65 years in 2020. | Community detection (LPA) |
| Poverty rate | Poverty rate per state in 2022. | Community detection (LPA) Random forest models  K-means model |
| State population | Population per state in 2022. | Community detection (LPA) |
| State population growth | State population growth since 2010. | Random forest models |
| Population density per county | Population density per county. Computed based on county population and county land area. | Random forest models  K-means model |
| Median age | Median age per county in 2019. | Random forest models  K-means model |
| Median household income | Median household income per county in 2019. | Random forest models |
| Median individual income | Median individual income per county in 2019. | Random forest models  K-means model |
| Unemployment rate | Unemployment rate among those aged 20-64 per county in 2019. | Random forest models |
| High school graduates | Percentage of those aged 25+ that are high school graduates per county in 2019. | Random forest models |
| Bachelor graduates | Percentage of those aged 25+ that earned a Bachelor’s degree or higher per county in 2019. | Random forest models |
| Persons per household | Average number of persons per household per county in 2019. | Random forest models |
| Households with smartphone | Percentage of households that have a smartphone per county in 2019. | Random forest models |
| Households with limited English | Percentage of households that speak limited English per county in 2019. | Random forest models |
| One-unit housing structures | Percentage of housing units in one-unit structures per county in 2019. | Random forest models |

1. We considered using measures about the racial composition of counties. However, we decided against it since we are interested in the explanatory factors behind the evolution of Covid-19. By itself, race is not an explanatory factor and would function as a potentially bias-generating proxy instead (see for example <https://www.educative.io/blog/racial-bias-machine-learning-algorithms>). Thus, we aimed to focus on less biased measures of social disadvantages. [↑](#footnote-ref-0)