A scalable pipeline for COVID-19: the case study of Germany, Czechia and Poland.

Wildan Abdussalam[1,*], Adam Mertel[1], Kai Fan[1], Lennart Schüler[1,2], Weronika Schlechte-Wełnicz[1] and Justin M. Calabrese[1,3,4]

[1]*Center for Advanced Systems Understanding, Helmholtz-Zentrum Dresden-Rossendorf, Untermarkt 20, 02826 Görlitz, Germany*

[2]*Department of Computational Hydrosystems, Helmholtz Centre for Environmental Research (UFZ), Permoserstraße 15, 04318 Leipzig, Germany*

[3]*Department of Ecological Modelling, Helmholtz Centre for Environmental Research (UFZ),Permoserstraße 15, 04318 Leipzig, Germany*

[4]*Department of Biology, University of Maryland, College Park MD, Maryland, USA*

Abstract

Throughout the coronavirus disease 2019 (COVID-19) pandemic, decision makers have relied on forecasting models to determine and implement non-pharmaceutical interventions (NPI). In building the forecasting models, continuously updated datasets from various stakeholders including developers, analysts, and testers are required to provide precise predictions. Here we report the design of a scalable pipeline which serves as a data synchronization to support inter-country top-down spatiotemporal observations and forecasting models of COVID-19, named the *where2test*, for Germany, Czechia and Poland. We have built an operational data store (ODS) using PostgreSQL to continuously consolidate datasets from multiple data sources, perform collaborative work, facilitate high performance data analysis, and trace changes. The ODS has been built not only to store the COVID-19 data from Germany, Czechia, and Poland but also other areas. Employing the dimensional fact model, a schema of metadata is capable of synchronizing the various structures of data from those regions, and is scalable to the entire world. Next, the ODS is populated using batch Extract, Transfer, and Load (ETL) jobs. The SQL queries are subsequently created to reduce the need for pre-processing data for users. The data can then support not only forecasting using a version-controlled Arima-Holt model and other analyses to support decision making, but also risk calculator and optimisation apps [1, 2]. The data synchronization runs at a daily interval, which is displayed at https://www.where2test.de.

1. Introduction

In building forecasting models of COVID-19, many researchers employ the training datasets provided by each country's representative institutions, e.g., Robert Koch Institute in Germany. The publicly accessible COVID-19 data provided in raw textual format, such as CSV, JSON, and XML are downloaded and analysed by the researchers employing either statistical or machine learning approaches. However, the data are unwell structured and require heavy pre-processing as well as ingestion activities for further analysis. This method is inherently inefficient due to identical and manual parallel pre-processing of the RKI data (using e.g. python or R scripts) performed by each researcher. This reduces the efficiency of each and everyone's work as all have to spend hours and days in pre-processing data before coming to modeling and forecasting. Advanced computing infrastructures and novel software pipelines are crucial tools to synchronize the data structures which originate from various sources and to extremely reduce heavy pre-processing [3]. They serve as essential prerequisites to realise the data surveillance and outbreak response management, which have been implemented in fighting other endemic diseases [4, 5, 6, 7].

To date, the data management have been applied in controlling the outbreak of COVID-19 [8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22]. Most of them provide maps and the prevalent data in the following regional level: (i) National level, e.g., COVID-19 data of World wide [10], Europe [11, 12, 13], and Latin America [14]; (ii) State and county levels, e.g., the COVID-19 data warehouse for Italy [15], COVID-19 dashboard for UK [17], the COVID-19 dashboard for Maryland [18], and for Germany [19].; (iii) County level, e.g., Dresden, Germany [20]. More completed version is provided by the John Hopkins University [21], which serves the dashboard and the prevalent data for each regional level in the USA as well as for most of countries around the world. Likewise, the similar method in the presence of semi-automatic validation strategy was conducted to check the data quality of daily updated numbers with governmental/official data sources [22]. However, most of dashboards and data warehouses have not provided the features to let the users perform an inter-country top-down spatiotemporal observation, i.e., observing the inter-country prevalence and simultaneously being able to observe to the microscopic level (nation → state → county → municipality). The features could provide insights, for example, to study COVID-19 border dynamics which have been so far attracted considerable atten-

tions [23, 24, 25, 26]. Moreover, they are lack of forecasting features, which play a key role in predicting the future prevalence as well as determining non pharmaceutical interventions (NPI). A tremendous number of forecasting models, e.g., agent-base [27], machine learning [28, 29], combination model [30, 31], compartment model [32, 33, 34, 35, 36], time series [37, 38, 39, 40, 41, 42, 43] have employed government datasets to provide essential inputs for public decisions. However, most of datasets that were used in those studies are limited to the specific time window which are likely to produce different results when the datasets are updated. Establishing a system of daily-updated-datasets assisted forecasts, therefore, is an alternative to improve their consistency and precision.

In this paper, we address the aforementioned issues by proposing the design of a scalable pipeline which allow us to perform the top-down spatiotemporal observation among Germany, Czechia, and Poland as well as to perform daily forecasts. The method of the pipeline which consists of extraction of various data sources and the ODS is described in subsec 2.1. More specifically, we will describe the dimensional fact database model and a daily migration process which underline the data synchronization between various data sources and our database server. We employ the dimensional fact model due to more flexibility and versatility in building spatiotemporal aggregation functions than the nanocubes model [44, 45]. Next, in subsec 2.2 we will describe the time-series forecasting models which are supported by the presence of the ODS. Furthermore, the automatic system of daily forecasts owing to the presence of the pipeline will be laid out in this sub section. In Sec. 3, we will describe facilities that have been established due to the presence of the ODS. In order to demonstrate the inter-country top-down spatiotemporal observations, an analysis will begin from the macroscopic scale in which the study of the virus spread across the national borders is described in subsec 3.1. Herein we consider the border among Germany, Czechia and Poland as a study case. In subsec 3.2, we explore more microscopic level by applying a daily-updated-datasets assisted forecast for the prevalence in the state of Saxony, Germany. Last but not least, in subsec 3.3, most microscopic level that we will demonstrate is a superspreading event at a slaughter house in Gütersloh, Lower Saxony, Germany. As the COVID-19 situation begins to enter an endemic phase, a study of superspreading event will provide essential information to trace the COVID-19 transmission after a mass event.
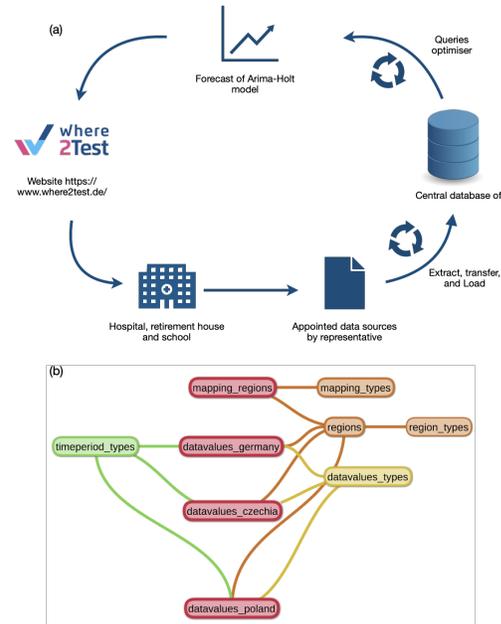
Figure 1: (a) A workflow of data pipeline Hospitals, retirement houses, and schools of Germany, Czechia and Poland update the data of COVID-19 cases, vaccines and tests to the representative government institutions. A daily automatic ETL step is performed to synchronize the data sources and central database of CASUS. A daily and weekly automatic forecast employing, e.g. Arima-Holt model, is applied to provide rapid predictions. The predictions and the actual data are shown in the where2test website; (b) The scalable dimensional fact model. Datavalues and datavalue types represent **measures**, while region types and timeperiod types represent spatial and temporal **dimensions**, respectively.

2. Methods

2.1. Data Pipeline

Fig. 1a shows a workflow of the data pipeline. The hospitals, retirement houses and schools register the daily number of the COVID-19 cases and vaccines to the representative government. In order to consolidate these data, the relational database is built based on dimensional fact model [46]. Having established the relational database, the daily automatic extract, transfer and load (ETL) step is performed to migrate and integrate the data sources to the PostgreSQL database of CASUS HZDR (see Suplementary materials 7.1). Next, we create SQL inquiries-based views to be analysed by our researchers using the forecast and machine learning methods. The tested and completed analysis methods are set in the master stage and the other tested methods are set in the

develop stage. Only the forecasting method in the master stage is integrated in the automatic pipeline.

The dimensional fact model is shown in Fig. 1b. The model consists of three main concepts: (i) **Facts**, that refer to a subject of study (e.g., the study of infected, dead, recovered, hospitalised, test and vaccinated cases due to COVID-19); (ii) **Measures**, that refer to the quantitative data of the concept (*i*). The measured data are stored in the table of datavalues. The tables of datavalues contain the number of infected, dead, recovered, hospitalised, test, and vaccinated cases due to COVID-19 in a given time and place. To date, the schema consists of three datavalues, i.e., datavalues of Germany, Czechia and Poland; (iii) **Dimensions**, that refer to temporal and spatial attributes. As the measured data are provided in a given time and place, the table of time period types and regions is necessary. The former stores the type of time period which consists of day and week data type; and the latter stores the necessary information of regions which consist of the name, abbreviation, ID of regions, ID of region type, geometry and population. The table of regions depends on the table of region types. The regions are categorised based on their sizes. The order of ascending sizes starts from municipality, county, state and nation. For Germany, the order of region type starts from Gemeinde, Kreise and Bundesland. Similar to Germany, Poland consist of Gmina, Powiat, and Wojewodztwo. Different from Germany and Poland, Czechia consist of 4 level, Obec, Orp, Okres and Kraj. The spatial and temporal attributes are connected by means of hierarchies to represent a -to-one relationship between them. The table of mapping_types contains the hierarchical type of the spatial attributes, e.g., for Germany (Gemeinde to Kreise, Kreise to Bundesland), for Czechia (Obec to Orp, Orp to Okres and Okres to Kraj), and Poland (Gmina to Powiat and Powiat to Wojewodztwo). Next, a many-to-one relationship between those spatial hierarchies are stored in the table of mapping_regions. Moreover, the table of timeperiod_types consists of the hierarchical type of the temporal attributes.

Aggregation functions are applicable on the measures along the temporal and spatial dimensions. For the former dimension, the weekly data are cumulative 7–day data. For example, a 7–day case reported on 13.03.2022 is an accumulation of the daily case for 07-13.03.2022. Moreover, for the latter dimension, county data are cumulative-municipality data. Not only accumulating the data from the municipality to a county level, in the presence of mapping regions table, it is possible to accumulate the data from the county to the state level as well as the state to the nation level. This allows us to scale the pipeline to other areas provided that the data of municipality are available from the sources.

2.2. Forecasts

We employ auto regression integrated moving average (ARIMA) and Holt's linear trend models to forecast the infected, test, and hospitalised data of COVID-19 for Saxony (Germany), Czechia, and Poland. The ARIMA model has been successfully employed in predicting other endemic diseases [47, 48, 49, 50]. The model features suitable prediction based on time analysis series which is capable of providing short horizon forecast for most COVID-19 cases around the world [38, 39, 40, 41, 42, 43]. To make the model consistent and avoid overfitting, the order parameter of the ARIMA model is fixed instead of using the auto ARIMA model. The ARIMA is improved by employing the Holt's linear trend model [51]. The Holt's model uses the exponential smoothing method to compute the weighted average of the past observation data [52]. The forecasts from the Holt's linear model have a trend, so the damped parameter is turned on to avoid this trend [53, 54, 52]. A self-defined mix function is used to compute the probability parameter m to combine the forecasts from two models and minimize the error. The Box-Cox transformation is used to normalize the input data [55, 52].

Our model provides a weekly forecast at first. In order to improve the daily variation and provide more real-time forecasts, we have built a daily forecast model. As the daily data have a clear weekly variation, the seasonal parameters are added to the model; and seasonal ARIMA (SARIMA) and Holt-Winters' seasonal model are employed for the daily forecasts [56, 51, 57]. Similar to the ARIMA model, the seasonal ARIMA model uses the fixed order and seasonal parameters. After comparing the errors from multiple methods, the additive method is selected for the Holt-Winters' seasonal model. The mix function is also used for the daily forecasts to combine the forecasts from two models and improve the forecasting accuracy. For study cases of (S)Arima-Holt model, in Sec. 3.2, we will provide the number of infections for Saxony, Germany. In addition to (S)ARIMA-Holt model, we employ outlier detection to identify and quantify Superspreading events. As suggested in [35], we identify and quantify superspreading events by using time series analysis based outlier detection methods. The rate of newly infected is modeled by an appropriate model, which could be something as simple as a rolling average to more elaborate ones as SIR-based models. The residues of the reported cases is used to identify outliers. At the same time, the residues can be used to quantify the size of a superspreading event.

3. Results

The presence of the pipeline has allowed us to provide following facilities: (i) The released data hub for dead and

infected cases of all counties and states in Germany [58], which allows a collaboration between CASUS research staffs and other external collaborators. The post-processing data serve as the clean data of daily infected and dead cases for county and state levels. In addition, we have also pre-processed the vaccination and hospitalization data for the county and municipal levels; (ii) The daily updated value of background risk for optimisation [1] and risk calculator apps [2], which defines the chance of an average person who lives in the focal area, and carries out daily activities, will be infected over a one week period; (iii) Blog posts which update current COVID-19 situations in Germany. An interesting example of the posts would be the relation between the vaccination rate and the 7-day incidence in all states of Germany [59]; (iv) Forecast- and model-based analysis. We explore the study cases mentioned in Sec. 1, and begin by investigating of the virus spread across the national borders of Germany, Czechia, and Poland.

3.1. Analysis of the virus spread across the national borders

COVID-19 spread among people. Therefore, human mobility is one of the most important factors defining the trend of spatiotemporal spreading of the virus. Understanding human mobility allows us to predict the spatiotemporal character of spread, evaluate the government steps restrictions, and provide effective non-pharmaceutical interventions. Primarily due to the heterogeneity of the sources and the interest scope of the particular research groups and communities, most of the COVID-19 research stays within the boundaries of one country. While most human mobility happens in the extent of one country or region, notably in Europe, the national border's mitigating effect is generally diminishing. To study the impact of the national border, several research papers [60, 61] applied various methodologies of geostatistics and geospatial modeling. More thorough quantification of the effect of border presence and international mobility on the epidemy requires a data storage integrating heterogeneous datasets across more countries.

The presented ODS infrastructure offers a possibility to study the spatiotemporal character of the virus spread on more levels, considering the effect of the national border. First, for our case study comprising the countries of Germany, Poland, and Czechia, we explored the correlation of new cases in the region, the distance and the border presence. We observed that the neighbour regions tend to have similar incidence values in the absence of barrier in the form of a national border among them. This step followed the research of McMahon et al. [62], which showed a strong spatial autocorrelation of incidence values in the USA.

Further, we calculated the average time-lagged pair-
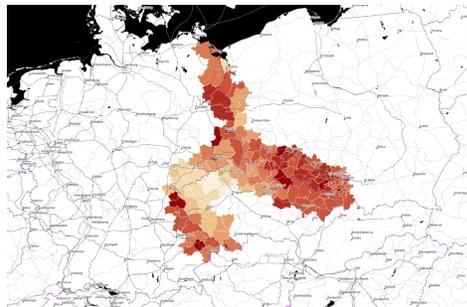
Figure 2: Difference in the pair-wise correlations for regions within a 100 kilometer radius inside and outside the country. The red color represents the regions with the strongest difference, indicating the spread of the virus across the national borders.

wise correlations for each region considering the regions in the radius of 100 kilometers, (i) within the same country, (ii) outside this country. The difference of these values can be seen in Fig. 2. The bigger difference represents regions where the incidence correlates much better than the regions within the same country, indicating a strong national border effect on the virus spread.

In the next step [63], we quantified the mitigation effect of the national border in more detail. We picked the state of Saxony in Germany and the neighboring regions in Czechia. For both countries, we collected and integrated the incidence data on the level of single municipalities. For each municipality, we constructed a local regression model which estimated the effect of three parameters, (i) border presence, (ii) municipality size, and (iii) temporal distance from other municipalities, on the spread of the virus. Based on this model, we identified very small-scale areas susceptible to a more intensive inter-national spread of the COVID-19.

The top-down approach we selected for the study on the national border effect is possible thanks to the scalability of the implemented dimensional-fact model. This principle allows the ODS to comprise various administrative levels and combine various relevant topics within the perspective of spacetime.

3.2. Weekly and daily forecast of Arima-Holt and Sarima-Holt

For the case study, we provide a short-time forecast of 7-day incidence up to 4 horizons performed on 13-04-2022 using Arima-Holt model for Saxony, Germany. We used a training dataset of 13-04-2022 version which consists of the historical weekly data of Saxony and its counties from 01-03-2020 to 10-04-2022. The weekly data are automated-daily-updated data which are aggregated on Sunday (see
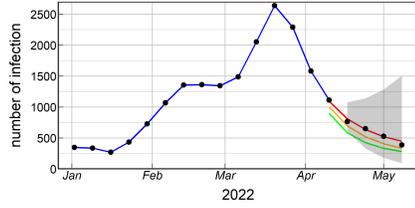
Figure 3: 7-day incidence of infected cases Jan - 8 May 2022 for Saxony, Germany. The black dots denote the historical data, the blue line (—) denotes a line guidance for the historical data, and the green (—), orange (—), and red line (—) denotes the result of forecast using the Arima-Holt model performed on 10-04-2022, 11-04-2022, and 13-04-2022, respectively. The grey area shows the lower and upper limits of the forecast for 13-04-2022.

Sec. 2.1). Although we update the data daily, for the case of Germany, the current and previous-day data are unavailable. In addition, the previous third day data are still to be updated from the source. When the forecast was performed on Sunday 10-04-2022, the number of infection on that day was less than the number of the same day for the following-day version. As a result, this produces inaccurate forecast (see Fig. 3). As the day elapsed, more cases were automatically added and aggregated to the last Sunday data. Consequently, the performed forecast on 13-04-2022 provides higher exponent than the one with the dataset version of 10 and 11-04-2022. Moreover, the dataset of Wednesday consists of relatively-stable version. Therefore, the forecast is performed every Wednesday due to the consistency of data source for the last Sunday.
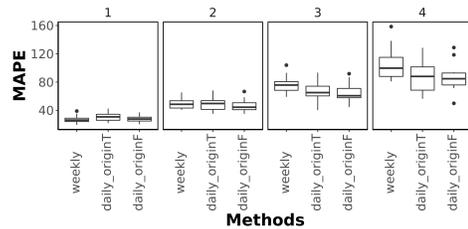
Figure 4: Mean absolute percentage error of Arima-Holt (weekly), Sarima-Holt in the presence of Box-cox transformation (daily_originT), and Sarima-Holt in the absence of Box-cox transformation (daily_originF) for 1th - 4th horizon.

In order to check the four-horizon forecast, we compare it to the weekly-historical data updated on 11-05-2022. The latter consists of relatively stable data from 17-04-2022 to 08-05-2022. As shown in Fig. 3, the weekly-historical data is surprisingly in quantitative agreement with the four-horizon forecast. However, this agreement occurs occasionally. When the forecast is performed in

a different day, a deviation from the actual data for the following 4 horizons is likely to occur. Additional realisations of Arima-Holt forecast in Saxony and its counties, therefore, were performed to improve statistics. The realisations were performed every Wednesday from 05-01-2022 to 18-05-2022 in which the version-control dataset were employed as training and test datasets. An example would be a realisation of the Forecast on 05-01-2022. We used the weekly data version of 05-01-2022 as its training dataset and the weekly data version of the following 1st, 2nd, 3rd and 4th week as its test datasets. For each region, we then recorded a deviation of the forecast result from the historical data and quantified it as mean absolute percentage error (MAPE). As shown in Fig. 4, the weekly Arima-Holt provides relatively low MAPE for the first and second horizon. For the third and fourth horizon, however, the range of MAPE tends to be wider than the first and second.

Therefore, we performed the Sarima-Holt model to improve the performance of forecast for the third and fourth horizon. Owing to daily-updated data, the version-control of daily data is employed as the seasonal parameters. In addition to the daily data, the Sarima-Holt forecast was performed using the same version-control weekly data employed to the Arima-Holt model. For the daily data, we removed the current and two previous-day data due to zero values for current and yesterday data, and inconsistent data for the previous third day. We then compared its performance in the presence and the absence of the Box-Cox transformation (BCT) used to normalize the input data. As shown in Fig. 4, the Sarima-Holt model in the absence of the BCT provides lower MAPE than either the Arima-Holt or the Sarima-Holt in the presence of the BCT for not only the first and second horizons, but also the third and four horizons.

3.3. Superspreading events

Superspreding events play an important role in the dispersion dynamics of COVID-19 [64]. However, one of the most commonly used epidemiological model types, the compartment models, are not able to accurately capture these events [35, 65]. We are currently working on a solution to the problem by using outlier detection methods on a county level. Many different methods exist and they can produce more robust results, when more than one timeseries is taken into account. A database as presented in this work is very advantageous, as it makes it very convenient to query the reported infections from all neighboring counties and use this additional data to more robustly identify outliers, which might be superspreading events. The largest confirmed superspreading to date in Germany with 1766 infections happened in a meat processing facility in the North Rhine-Westphalian district of Gütersloh in June 2020. The facilities' environmental

conditions combined with relatively close physical distance between workers were likely the main reason for efficient aerosol transmission [66]. We take this event as an example to show the result of a Z-score based outlier detection method (Fig. 5).
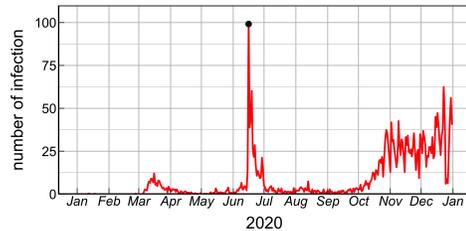
Figure 5: The official reported COVID-19 daily incidence per 100.000 inhabitants in the district of Gütersloh. A superspreading event in a meat processing plant in June 2020 is successfully identified by an outlier detection method based on the Z-score (the black dot).

4. Discussions

Our analysis, implementing the pipeline in the presence of dimensional fact model has allowed us to daily migrate the data efficiently due to the functions of spatiotemporal aggregation. To provide the weekly data of counties, states, and nations, we only migrate the data of daily municipalities/counties (depends on the data availability of each nation) to the database server which are then aggregated to the higher spatiotemporal level. This model provides more advantages than the nanocubes model [44, 45]. For the nanocubes model, each spatial (municipality, county, state and nation) and temporal (daily and weekly) data are required to be migrated to the database server. Consequently, this leads to a longer migration process than the one performed using the dimensional fact model. Moreover, its spatiotemporal mapping enables us to perform an efficient table join among national data which is confirmed by the application on the Subsec. 3.1.

The presence of daily-updated data due to the presence of the pipeline has allowed us to develop the Sarima-Holt model. The model shows more robust prediction for longer horizon than the Arima-Holt one. More specifically, the Sarima-Holt in the absence of the BCT outperforms the Arima-Holt model for the third and fourth horizon. This performance is due to the seasonal-parameter contribution to the model. As a result, the forecast tends to better predict for the third and fourth horizon. In contradiction, the Sarima-Holt in the presence of the BCT provides lower performance than the absence one due to less variation of the training data after BCT (see Fig. 7).

The Sarima-Holt model is trained by the daily data, and the variation of the data could make the model more sensitive to the infection change compared to the Arima-Holt model trained by the weekly data. However, the BCT reduces the variation of the daily data, and consequently the daily forecasts perform worse than in the absence of the BCT.

5. Conclusion

Our work has demonstrated the utility of the data pipeline for top-down spatiotemporal analysis. We have first shown the macroscopic analysis, in which the investigation of the virus spread across the national border is presented. At more microscopic level, we have demonstrated data-driven approach due to the presence of the pipeline which is applied to the prevalence of the county region. The daily-updated data has improved the precision of the model for longer horizon. This data-driven epidemic models provide more realistic forecast results than either the parsimonious [34] or more number of parameters with agent-based method [27] due to the usage of daily-updated data. This may contribute to public health policy making, including contributing to public health forecasting teams. Last but not least, exploring to lower level of region, we have demonstrated that the outlier model is applicable to capture the superspreading event which occurred in 2020. These have shown that our work is capable of performing top-down analysis as well as rapid and precise forecasts due to the presence of the pipeline.
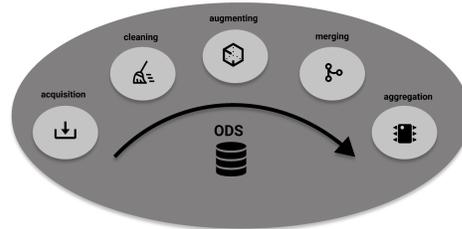
6. Data sources

- COVID-19 data for Germany, Czechia and Poland.
  - Robert Koch Institute
  - Czech Ministry of Health
  - Polish Ministry of Health
  - Age-based hospitalisation of state level for Germany (https://github.com/KITmetricslab/hospitalization-nowcast-hub/blob/main/data-truth/COVID-19/).
  - Age-based and type-based doses of vaccine for county level (https://github.com/robert-koch-institut/COVID-19-Impfungen_in_Deutschland/blob/master/Aktuell_Deutschland_Landkreise_COVID-19-Impfungen.csv).
  - COVID-19 infected, recovered, hospitalised and dead cases of Dresden (http://daten.dresden.de/duva2ckan/files/de-sn-dresden-corona_-_covid-19_-_fallzahlen_md1_dresden_2020ff/content).

- COVID-19 infected, dead, and test cases of Czechia for Municipality level (https://onemocneni-aktualne.mzcr.cz/api/v2/covid-19/).
- Age-based and gender-based infected and dead cases for county level of Germany (https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4).
- COVID-19 cases for municipality level of Saxony, Germany (https://www.coronavirus.sachsen.de/corona-statistics/rest/infectionOverview.jsp).
- COVID-19 cases for county level of Saxony, Germany (https://media.githubusercontent.com/media/robert-koch-institut/SARS-CoV-2_Infektionen_in_Deutschland/master/Aktuell_Deutschland_SarsCov2_Infektionen.csv)
- COVID-19 infected, dead, and test cases for county level of Poland (https://wojewodztwa-rcb-gis.hub.arcgis.com/pages/dane-do-pobrania).
- COVID-19 vaccine for county level of Poland (https://www.gov.pl/web/szczepimysie/raport-szczepien-przeciwko-covid-19).
- COVID-19 types in Sachsen (https://www.coronavirus.sachsen.de/infektionsfaelle-in-sachsen-4151.html).

- Dictionaries of regions.
  - Administrative areas in Germany (https://gdz.bkg.bund.de/index.php/default/digitale-geodaten/verwaltungsgebiete.html).
  - Administrative areas in Poland (https://gis-support.pl/baza-wiedzy-2/dane-do-pobrania/granice-administracyjne/)
  - Administrative areas in Czechia (https://geoportal.cuzk.cz/(S(1nhx02lray0vkrhce1y2d53d))/Default.aspx?mode=TextMeta&text=dSady_RUIAN&side=dSady_RUIAN)
  - Population numbers in Czech municipalities (https://www.czso.cz/csu/czso/pocet-obyvatel-v-obcich-k-112021)
  - Postal codes in Germany (https://www.geonames.org/postal-codes/postleitzahlen-deutschland.html)
  - Population numbers in Poland (https://stat.gov.pl/obszary-tematyczne/ludnosc/ludnosc/ludnosc-stan-i-struktura-ludnosci-oraz-ruch-naturalny-w-przekroju-terytorialnym-stan-w-dniu-30-06-2021,6,30.html)

7. Supplementary information

7.1. Data workflow

We use https://www.talend.com/products/talend-open-studio/ to perform data migration. The migration between the data sources and the PostgreSQL database of CASUS HZDR has been performed as follows:

Figure 6: Data workflow of the ETL process (see texts for its description).

1. **Data acquisition**
   The data are automatically downloaded from sources 6. They are subsequently stored on the repository of where2test server. The downloaded data serve as data inputs of a migration process.

2. **Dictionaries and data augmentation**
   To integrate and further augment data from heterogeneous sources (various forms, schema, temporal and spatial extent), we needed to prepare a list of dictionaries. We formed a dictionary for each spatial level in every country to cover all regions in our datasets. Here we included the unique region id, all alternative names, full names, geometries, and population numbers. This concept can be further extended to other values such as socioeconomical parameters, and information about the region. This way we are able to maintain the consistency in all datasets and enable their integration process. The list of sources used for building the dictionaries can be found in section Data Sources 6.

3. **Data cleaning**
   We migrate first **timeperiod_types**, **region_types**, **datavalues_types**, and **mapping_types**. While migrating the data to those tables, primary key are automatically set by a transformator (The script which migrates the data to the postgreSQL database.). Next, the primary key of those tables serves as the foreign key of other tables following the table relation shown in Fig. 1b. An example would be a table of **regions** which contains intrinsic ID set by representative governments. In order to differentiate ID among Germany, Czechia and

Poland, we add 'DE', 'CZ', 'PL', respectively, followed by the intrinsic ID. For the table of **regions**, the primary key of **region_types** serves as its foreign key. The intrinsic IDs are categorised based on the ID of region types. A specific example would be Dresden, whose the intrinsic ID 14162. After cleaning processes, the intrinsic ID will be DE 14162 and categorised to the state level of *Kreise*.

Having migrated the data to the aforementioned tables, the table of **mapping_regions** is occupied by the spatial-relation data. It contains the foreign key of the mapping type ID. An example would be a county Dresden. Dresden are mapped onto the state of Saxony and categorized to the mapping type **Kreis_To_Bundesland**. Next, the table of datavalues for nations is occupied by the data input. The datavalues table consists of three foreign keys which originate from the tables of **timeperiod_types**, **regions**, **datavalues_types**. In the presence of these foreign keys, a data merging process is feasible, which is described on the following item.

4. **Data merging** In addition to the aforementioned three-foreign keys, date is set as the fourth attribute which allow us to perform data merging through inner join of tables. The inner join is employed to cleanly merge and avoid duplicated data on the table of datavalues. For instance, daily infected data of the lowest-level region for period of date are migrated to the table of **datavalues_germany**. When the data sources are updated, they sometimes update the cases of the elapsed date. Inner join method allows us to automatically update the value of the elapsed date by the latest value. Moreover, when the new data with the latest date are present from the source, it allows automatic addition of the data to the table.

5. **Data aggregation** The presence of daily data of the lowest regions allow us to perform both time and spatial aggregations. Using functions, the time aggregation from daily to weekly period is feasible. Moreover, as mentioned on the Sec. 2, the spatial aggregation from the low to the high region level is allowable in the presence of the **mapping_regions** table.

7.2. Additional forecasting results

Acknowledgments
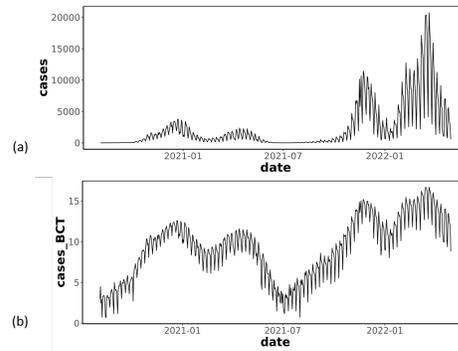
Figure 7: Time series of daily infected cases from Aug. 5, 2020 to Apr. 30, 2022 (a) before and (b) after Box-Cox transformation, respectively.

References

[1] M. Davoodi, A. Batista, A. Senapati, W. Schlechte-Welnicz, B. Wagner, J. M. Calabrese, Modeling COVID-19 optimal testing strategies in long-term care facilities: An optimization-based approach, arXiv (2022). URL: https://arxiv.org/abs/2204.02062. doi:10.48550/ARXIV.2204.02062.

[2] M. Davoodi, A. Senapati, A. Mertel, W. Schlechte-Welnicz, J. M. Calabrese, Optimal Workplace Occupancy Strategies during the COVID-19 Pandemic, arXiv (2022). URL: https://arxiv.org/abs/2204.01444. doi:10.48550/ARXIV.2204.01444.

[3] J. L. Raisaro, others, SCOR: A secure international informatics infrastructure to investigate COVID-19, Journal of the American Medical Informatics Association 27 (2020) 1721–1726. doi:10.1093/jamia/ocaa172.

[4] A. v. Wangenheim, A. Savaris, A. F. Borgatto, A. d. S. Inácio, Integrating Online Georeferenced Epidemiological Analysis and Visualization into a Telemedicine Infrastructure – First Results, medRxiv (2019). URL: https://www.medrxiv.org/content/10.1101/19000554v1.full.

[5] C. Fähnrich, others, Surveillance and Outbreak Response Management System (SORMAS) to support the control of the Ebola virus disease outbreak in West Africa, Euro Surveill 20 (2015) 21071. doi:https://doi.org/10.2807/1560-7917.es2015.20.12.21071.

[6] R. N. Smith, others, InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data, Bioinformatics 28 (2012) 3163–3165. doi:https://doi.org/10.1093/bioinformatics/bts577.

[7] C. Pfander, B. Anar, F. Schwach, T. D. Otto, M. Brochet, K. Volkmann, M. A. Quail, A. Pain, B. Rosen, W. Skarnes, J. C. Rayner, O. Billker, A scalable pipeline for highly effective genetic modification of a malaria parasite, Nature Methods 8 (2011) 1078–1082. URL: https://doi.org/10.1038/nmeth.1742. doi:10.1038/nmeth.1742.

[8] P. Kostkova, others, Data and Digital Solutions to Support Surveillance Strategies in the Context of the COVID-19 Pandemic, Frontiers in Digital Health 3 (2021). doi:https://doi.org/10.3389/fdgth.2021.707902.

[9] J. Budd, others, Digital technologies in the public-health response to COVID-19, Nature medicine 26 (2020) 1183–1192. URL: https://www.nature.com/articles/s41591-020-1011-4.

[10] F. A. Binti Hamzah, C. Hau, H. Nazri, D. Ligot, G. Lee, M. Shaib, U. Zaidon, A. Abdullah, M. H. Chung, C. Ong, P. Chew, CoronaTracker: Worldwide COVID-19 Outbreak Data Analysis and Prediction (2020). doi:10.2471/BLT.20.255695.

[11] E. Centre, European Centre for Disease Prevention and Control, 2022. URL: https://qap.ecdc.europa.eu/public/extensions/covid-19/covid-19.html#global-overview-tab.

[12] A. Naqvi, COVID-19 European regional tracker, Scientific Data 8 (2021) 181. URL: https://doi.org/10.1038/s41597-021-00950-7. doi:10.1038/s41597-021-00950-7.

[13] c. eudata, covid19-eu-data, 2020. URL: https://github.com/covid19-eu-zh/covid19-eu-data.

[14] c.-. latinoamerica, Latin America Covid-19 Data Repository by DSRP, 2020. URL: https://github.com/DataScienceResearchPeru/covid-19_latinoamerica.

[15] G. Agapito, C. Zucco, M. Cannataro, COVID-WAREHOUSE: A Data Warehouse of Italian COVID-19, Pollution, and Climate Data, Environmental Research and Public Health 17 (2020). doi:https://doi.org/10.3390/ijerph17155596.

[16] R. K. Arora, A. Joseph, J. Van Wyk, S. Rocco, A. Atmaja, E. May, T. Yan, N. Bobrovitz, J. Chevrier, M. P. Cheng, T. Williamson, D. L. Buckeridge, SeroTracker: a global SARS-CoV-2 seroprevalence dashboard, The Lancet Infectious Diseases 21 (2021) e75–e76. URL: https://doi.org/10.1016/S1473-3099(20)30631-9. doi:10.1016/S1473-3099(20)30631-9, publisher: Elsevier.

[17] c. GovUK, Interactive map of cases, 2022. URL: https://coronavirus.data.gov.uk/details/interactive-map/cases.

[18] g. maryland, Coronavirus Disease 2019 (COVID-19) Outbreak, 2022. URL: https://coronavirus.maryland.gov.

[19] c. d. rki, Robert Koch-Institut: COVID-19-Dashboard, 2022. URL: https://experience.arcgis.com/experience/478220a4c454480e823b17327b2bf1d4/page/Landkreise/.

[20] c. dresden, Corona-Dashboard Dresden, 2022. URL: https://experience.arcgis.com/experience/d2386f3214c1451c81b242be69bb3d50.

[21] E. Dong, H. Du, L. Gardner, An interactive web-based dashboard to track COVID-19 in real time, The Lancet Infectious Diseases 20 (2020) 533–534. URL: https://doi.org/10.1016/S1473-3099(20)30120-1. doi:10.1016/S1473-3099(20)30120-1, publisher: Elsevier.

[22] D. Sha, Y. Liu, Q. Liu, Y. Li, Y. Tian, F. Beaini, C. Zhong, T. Hu, Z. Wang, H. Lan, Y. Zhou, Z. Zhang, C. Yang, A spatiotemporal data collection of viral cases for COVID-19 rapid response, Big Earth Data 5 (2021) 90–111. URL: https://doi.org/10.1080/20964471.2020.1844934. doi:10.1080/20964471.2020.1844934, publisher: Taylor & Francis.

[23] Han Xiaoyi, Xu Yilan, Fan Linlin, Huang Yi, Xu Minhong, Gao Song, Quantifying COVID-19 importation risk in a dynamic network of domestic cities and international countries, Proceedings of the National Academy of Sciences 118 (2021) e2100201118. URL: https://doi.org/10.1073/pnas.2100201118. doi:10.1073/pnas.2100201118, publisher: Proceedings of the National Academy of Sciences.

[24] D. Laroze, E. Neumayer, T. Plümper, COVID-19 does not stop at open borders: Spatial contagion among local authority districts during England's first wave, Social Science & Medicine 270 (2021) 113655. URL: https://www.sciencedirect.com/science/article/pii/S0277953620308741. doi:10.1016/j.socscimed.2020.113655.

[25] M. Grimée, M. Bekker-Nielsen Dunbar, F. Hofmann, L. Held, Modelling the effect of a border closure between Switzerland and Italy on the spatiotemporal spread of COVID-19 in Switzerland, Spatial Statistics (2021) 100552. URL: https://www.sciencedirect.com/science/article/pii/S2211675321000622. doi:10.1016/j.spasta.2021.100552.

[26] M. P. Hossain, A. Junus, X. Zhu, P. Jia, T.-H. Wen, D. Pfeiffer, H.-Y. Yuan, The effects of border control and quarantine measures on the spread of COVID-19, Epidemics 32 (2020) 100397. URL: https://www.sciencedirect.com/science/article/pii/S1755436520300244. doi:10.1016/j.epidem.2020.100397.

[27] Q.-H. Liu, others, Model-based evaluation of alternative reactive class closure strategies against

COVID-19, Nat. Com. 13 (2022). doi:10.1038/s41467-021-27939-5.

[28] H. Bastani, others, Efficient and targeted COVID-19 border testing via reinforcement learning, Nature 559 (2021). URL: https://www.nature.com/articles/s41586-021-04014-z.

[29] S. Flaxman, others, Estimating the effects of non-pharmaceutical interventions on COVID-19 in Europe, Nature 584 (2020) 257. URL: https://www.nature.com/articles/s41586-020-2405-7.

[30] N. Haug, others, Ranking the effectiveness of worldwide COVID-19 government interventions, Nature Human Behaviour 4 (2020) 1303–1312. URL: https://www.nature.com/articles/s41562-020-01009-0.

[31] A. Liu, L. Vici, V. Ramos, S. Giannoni, A. Blake, Visitor arrivals forecasts amid COVID-19: A perspective from the Europe team, Annals of Tourism Research 88 (2021) 103182. URL: https://www.sciencedirect.com/science/article/pii/S016073832100044X. doi:10.1016/j.annals.2021.103182.

[32] S. Lai, others, Effect of non-pharmaceutical interventions to contain COVID-19 in China, Nature 585 (2020) 410. URL: https://www.nature.com/articles/s41586-020-2293-x.

[33] D. Fanelli, F. Piazza, Analysis and forecast of COVID-19 spreading in China, Italy and France, Chaos, Solitons & Fractals 134 (2020) 109761. URL: https://www.sciencedirect.com/science/article/pii/S0960077920301636. doi:10.1016/j.chaos.2020.109761.

[34] Bertozzi Andrea L., Franco Elisa, Mohler George, Short Martin B., Sledge Daniel, The challenges of modeling and forecasting the spread of COVID-19, Proceedings of the National Academy of Sciences 117 (2020) 16732–16738. URL: https://doi.org/10.1073/pnas.2006520117. doi:10.1073/pnas.2006520117, publisher: Proceedings of the National Academy of Sciences.

[35] L. Schüler, J. M. Calabrese, S. Attinger, Data driven high resolution modeling and spatial analyses of the COVID-19 pandemic in Germany, PLOS ONE 16 (2021) e0254660. URL: https://doi.org/10.1371/journal.pone.0254660. doi:10.1371/journal.pone.0254660, publisher: Public Library of Science.

[36] I. Rahimi, F. Chen, A. H. Gandomi, A review on COVID-19 forecasting models, Neural Computing and Applications (2021). URL: https://doi.org/10.1007/s00521-020-05626-8. doi:10.1007/s00521-020-05626-8.

[37] R. Salgotra, M. Gandomi, A. H. Gandomi, Time Series Analysis and Forecast of the COVID-19 Pandemic in India using Genetic Programming, Chaos, Solitons & Fractals 138 (2020) 109945. URL: https://www.sciencedirect.com/science/article/pii/S0960077920303441. doi:10.1016/j.chaos.2020.109945.

[38] S. Roy, G. S. Bhunia, P. K. Shit, Spatial prediction of COVID-19 epidemic using ARIMA techniques in India, Modeling Earth Systems and Environment 7 (2021) 1385–1391. URL: https://doi.org/10.1007/s40808-020-00890-y. doi:10.1007/s40808-020-00890-y.

[39] M.-J. Geng, H.-Y. Zhang, L.-J. Yu, C.-L. Lv, T. Wang, T.-L. Che, Q. Xu, B.-G. Jiang, J.-J. Chen, S. I. Hay, Z.-J. Li, G. F. Gao, L.-P. Wang, Y. Yang, L.-Q. Fang, W. Liu, Changes in notifiable infectious disease incidence in China during the COVID-19 pandemic, Nature Communications 12 (2021) 6923. URL: https://doi.org/10.1038/s41467-021-27292-7. doi:10.1038/s41467-021-27292-7.

[40] Y. Wang, C. Xu, S. Yao, L. Wang, Y. Zhao, J. Ren, Y. Li, Estimating the COVID-19 prevalence and mortality using a novel data-driven hybrid model based on ensemble empirical mode decomposition, Scientific Reports 11 (2021) 21413. URL: https://doi.org/10.1038/s41598-021-00948-6. doi:10.1038/s41598-021-00948-6.

[41] V. K. Sharma, U. Nigam, Modeling and Forecasting of COVID-19 Growth Curve in India, Transactions of the Indian National Academy of Engineering 5 (2020) 697–710. URL: https://doi.org/10.1007/s41403-020-00165-z. doi:10.1007/s41403-020-00165-z.

[42] A. K. Sahai, N. Rath, V. Sood, M. P. Singh, ARIMA modelling & forecasting of COVID-19 in top five affected countries, Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14 (2020) 1419–1427. URL: https://www.sciencedirect.com/science/article/pii/S1871402120302903. doi:10.1016/j.dsx.2020.07.042.

[43] D. Benvenuto, M. Giovanetti, L. Vassallo, S. Angeletti, M. Ciccozzi, Application of the ARIMA model on the COVID-2019 epidemic dataset, Data in Brief 29 (2020) 105340. URL: https://www.sciencedirect.com/science/article/pii/S2352340920302341. doi:10.1016/j.dib.2020.105340.

[44] L. Lins, J. T. Klosowski, C. Scheidegger, Nanocubes for Real-Time Exploration of Spatiotemporal Datasets, IEEE Transactions on Visualization and Computer Graphics 19 (2013) 2456–2465. doi:10.1109/TVCG.2013.179.

[45] A. Bosworth, J. Gray, A. Layman, H. Pirahesh, Data Cube: A Relational Aggregation Operator Generalizing Group-By, Cross-Tab, and Sub-Totals, Technical Report MSR-TR-95-22, Institute of Electrical and Electronics Engineers, Inc., 1995. URL: https://www.microsoft.com/en-us/research/publication/data-cube-a-relational-aggregation-operator-generalizing-group-by-cross-tab-and-sub-totals/.

[46] M. Golfarelli, D. Mario, S. Rizzi, The dimensional

fact model: a conceptual model for data warehouses, International Journal of Cooperative Information Systems 7 (1998) 215–247. doi:https://doi.org/10.1142/S0218843098000118.

[47] E. O. Nsoesie, O. Oladeji, A. S. A. Abah, M. L. Ndeffo-Mbah, Forecasting influenza-like illness trends in Cameroon using Google Search Data, Scientific Reports 11 (2021) 6713. URL: https://doi.org/10.1038/s41598-021-85987-9. doi:10.1038/s41598-021-85987-9.

[48] Y. Chen, Y. Zhang, Z. Xu, X. Wang, J. Lu, W. Hu, Avian Influenza A (H7N9) and related Internet search query data in China, Scientific Reports 9 (2019) 10434. URL: https://doi.org/10.1038/s41598-019-46898-y. doi:10.1038/s41598-019-46898-y.

[49] Z. He, H. Tao, Epidemiology and ARIMA model of positive-rate of influenza viruses among children in Wuhan, China: A nine-year retrospective study, International Journal of Infectious Diseases 74 (2018) 61–70. URL: https://www.sciencedirect.com/science/article/pii/S1201971218344618. doi:10.1016/j.ijid.2018.07.003.

[50] Q. Zeng, D. Li, G. Huang, J. Xia, X. Wang, Y. Zhang, W. Tang, H. Zhou, Time series analysis of temporal trends in the pertussis incidence in Mainland China from 2005 to 2016, Scientific Reports 6 (2016) 32367. URL: https://doi.org/10.1038/srep32367. doi:10.1038/srep32367.

[51] C. C. Holt, Forecasting seasonals and trends by exponentially weighted moving averages, International Journal of Forecasting 20 (2004) 5–10. URL: https://www.sciencedirect.com/science/article/pii/S0169207003001134. doi:10.1016/j.ijforecast.2003.09.015.

[52] R. J. Hyndman, G. Athanasopoulos, Forecasting: Principles and Practice., OTexts, 2018. URL: https://otexts.com/fpp2/.

[53] E. S. Gardner, E. McKenzie, Why the damped trend works, Journal of the Operational Research Society 62 (2011) 1177–1180. URL: https://doi.org/10.1057/jors.2010.37. doi:10.1057/jors.2010.37, publisher: Taylor & Francis.

[54] E. S. Gardner, E. Mckenzie, Forecasting Trends in Time Series, Management Science 31 (1985) 1237–1246. URL: https://doi.org/10.1287/mnsc.31.10.1237. doi:10.1287/mnsc.31.10.1237, publisher: INFORMS.

[55] V. M. Guerrero, Time-series analysis supported by power transformations, Journal of Forecasting 12 (1993) 37–48. URL: https://doi.org/10.1002/for.3980120104. doi:10.1002/for.3980120104, publisher: John Wiley & Sons, Ltd.

[56] S. C. Hillmer, G. C. Tiao, An ARIMA-Model-Based Approach to Seasonal Adjustment, Journal of the American Statistical Association 77 (1982) 63–70. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1982.10477767. doi:10.1080/01621459.1982.10477767, publisher: Taylor & Francis.

[57] P. R. Winters, Forecasting Sales by Exponentially Weighted Moving Averages, Management Science 6 (1960) 324–342. URL: https://doi.org/10.1287/mnsc.6.3.324. doi:10.1287/mnsc.6.3.324, publisher: INFORMS.

[58] W. Abdussalam, Post-processing data of daily dead and infected COVID-19 in Germany (2022). URL: https://zenodo.org/badge/latestdoi/462876343. doi:DOI:10.5281/zenodo.6336637.

[59] A. Mertel, M. Laqua, Where2Test visualization highlights strong link between pace of vaccinations and incidences, 2022. URL: https://www.where2test.de/blog#vaccination-maps.

[60] M. Eckardt, K. Kappner, N. Wolf, Covid-19 across european regions: The role of border controls (2020).

[61] M. Grimée, M. B.-N. Dunbar, F. Hofmann, L. Held, et al., Modelling the effect of a border closure between switzerland and italy on the spatiotemporal spread of covid-19 in switzerland, Spatial statistics (2021) 100552.

[62] T. McMahon, A. Chan, S. Havlin, L. K. Gallos, Spatial correlations in geographical spreading of covid-19 in the united states, Scientific Reports 12 (2022) 1–10.

[63] A. Mertel, J. Vyskočil, L. Schüler, W. Schlechte-Wełnicz, J. M. Calabrese, Fine-scale variation in the effect of national border on covid-19 spread: A case study of the saxon-czech border region, medRxiv (2022).

[64] J. E. Lemieux, K. J. Siddle, B. M. Shaw, C. Loreth, S. F. Schaffner, A. Gladden-Young, G. Adams, T. Fink, C. H. Tomkins-Tinch, L. A. Krasilnikova, K. C. DeRuff, M. Rudy, M. R. Bauer, K. A. Lagerborg, E. Normandin, S. B. Chapman, S. K. Reilly, M. N. Anahtar, A. E. Lin, A. Carter, C. Myhrvold, M. E. Kemball, S. Chaluvadi, C. Cusick, K. Flowers, A. Neumann, F. Cerrato, M. Farhat, D. Slater, J. B. Harris, J. Branda, D. Hooper, J. M. Gaeta, T. P. Baggett, J. O'Connell, A. Gnirke, T. D. Lieberman, A. Philippakis, M. Burns, C. M. Brown, J. Luban, E. T. Ryan, S. E. Turbett, R. C. LaRocque, W. P. Hanage, G. R. Gallagher, L. C. Madoff, S. Smole, V. M. Pierce, E. Rosenberg, P. C. Sabeti, D. J. Park, B. L. MacInnis, Phylogenetic analysis of SARS-CoV-2 in the Boston area highlights the role of recurrent importation and superspreading events, preprint, Epidemiology, 2020. doi:10.1101/2020.08.23.20178236.

[65] G. B. Libotte, L. dos Anjos, R. C. Almeida, S. M. C. Malta, R. S. Silva, Framework for enhancing the estimation of model parameters for data with a high level of uncertainty, preprint, Epidemiology, 2020.

URL: http://medrxiv.org/lookup/doi/10.1101/2020.12.17.20248389. doi:10.1101/2020.12.17.20248389.

[66] T. Günther, M. Czech-Sioli, D. Indenbirken, A. Robitaille, P. Tenhaken, M. Exner, M. Ottinger, N. Fischer, A. Grundhoff, M. M. Brinkmann, SARS-CoV-2 outbreak investigation in a German meat processing plant, EMBO Molecular Medicine 12 (2020) e13296. URL: https://doi.org/10.15252/emmm.202013296. doi:10.15252/emmm.202013296, publisher: John Wiley & Sons, Ltd.