Folkhalsomyndigheten's daily data download

Damian Ke

Contents

1 Background

During pandemics such as COVID-19 it is important to have correct type of data as it can be used for decision making in capacity planning or resource management (Klappe et al. 2022). The healthcare data from can be unstructured and inconsistent which can lead to incorrect conclusions (Klappe et al. 2022). In Sweden, Folkhälsomyndigheten (FHM) has the national responsibility of public health ("Folkhälsomyndigheten," n.d.). Lately, it has once again changed the way it provides COVID19 data concerning Sweden. The current way of providing data by FHM contain redundant data. Additionally, FHM periodically updates and potentially backcorrects its data regarding COVID-19 statistics.

1.1 Data Variations

The data from FHM have notable differences over time. Table 1, shows when different data sheets started to appear in the excel files.

Table 1: Data Variations in Excel

Data_Sheet	Introduced
Veckodata Region	2020-06-12
Veckodata Kommun_stadsdel	2020-06-12
Veckodata Riket	2021-01-23

Table 2, shows which information cannot be found in the excel files with comparison to the text files. The dataset included weekly data fields like 'ycov19ivavald', 'ycov19ivavkon', and 'dcov19ald', while the Excel files present this data in a cumulative format.

Table 2: Missing information in Excel Files

File	Missing_Information
ccov19Kon	Fall efter kön, region och vecka (tidsserie)
dcov19ald	Fall efter åldersgrupp, vecka och år, (Excel data is aggregated)
ecov19sabo	Fall bland personer 65 år och äldre med insats enligt socialtjänstlagen efter region och vecka (tidsserie)
ecov19sabosasong	Fall bland personer 65 år och äldre med insats enligt socialtjänstlagen
PCRtestVAr_k	Testade individer med PCR
PCRtestVAr_m	Testade individer med PCR
PCRtestVAr_s	Testade individer med PCR
ycov19ivavald	Intensivvårdade och avlidna fall efter åldersgrupp och vecka (tidsserie), (Excel data is aggregated)
ycov19ivavkon	Intensivvårdade och avlidna fall efter kön och vecka (tidsserie), (Excel data is aggregated)

2 Problem

In the data from FHM, it has been identified instances of redundancy. For effective analysis, the data needs to be consistent and useful. Therefore, the aim of the project is to implement a (preferably R) method that combines data from FHM's three data structures into a single database (preferably saved as an R file, or collection of csv files). One part is to identify the non-redundant information to save. Additionally, the data should be updated on consistent basis in the case of backcorrection. Lastly, the final goal of the project is to apply data mining techniques to explore whether different municipalities can be grouped into clusters based on the research question: "How similar are different municipalities in reporting the COVID-19 cases in Sweden?"

3 Result

This project has gone through several processes. Initially it went through data management to analyze the data and get the understanding of the data in hand. Thereafter, Extract, Transform and Load (ETL) process was implemented to store the correct type of data in the database. Lastly, the cleaned data has been used in data mining to answer the research question.

3.1 Data Management

The redundant information is for instance, the rows that include 'tot_antal', which represents summed-up data, and 'per 10000 inv' (or similar metrics), indicating the amount per 10,000 inhabitants. Considering redundancy and file size, it may be more efficient to exclude detailed values and instead use a separate reference table for things like population numbers. However, this approach has its trade-offs. Population figures can vary, leading to slight inaccuracies in short-term data due to changes in population sizes. Yet, the advantage is a more compact data file, which is easier to handle and analyze.

As Folkhälsomyndigheten periodically backcorrects data values. It's crucial to store both the original and revised data to maintain the correct data. To manage the growing size of the database efficiently, the proposed solution involves storing only the latest dataset along with a 'difference file' that records all backcorrections. This approach reduces the overall database size and simplifies analysis if the backcorrected data is of interest. However, reconstructing the original dataset from these difference files would require additional coding efforts.

Lastly, there are values as ".", "..", which may not bring any value for the analysis or usage. These values were interpreted as unknown data for years prior to 2020 or as placeholders for future uncollected data.

3.2 Building the Database

The data in this project results in a database as a collection of csv files. As there are multiple data sources, the solution is inspired by structure of "Data Warehouse". According to (Elmasri and Navathe 2015), Data warehouse is characterized by a warehouse with data that stores multiple data sources utilizing the ETL process. Although, in the "Fundamentals of Database Systems" (Elmasri and Navathe 2015), the data in Data Warehouse is usually stored as a multidimensional model. In the collected datasets, the data differs and cannot be captured in one single multidimensional model due to the nature of the collected data. For example, there are no values for number of intensive care and number of deaths of genders per regions. This can be hard to convert into a multidimensional model. Instead, the final database design of the csv files has been discussed with the researcher. Which will result in similar format as the txt files with some simplifications. For example, it won't contain per 10 000 rows but instead it can be calculated using the Municipalities_2022.csv. Names of regions were set to same name, which will make it easier to merge different tables together. Although, the ETL process will be followed.

3.2.1 ETL Process Overview

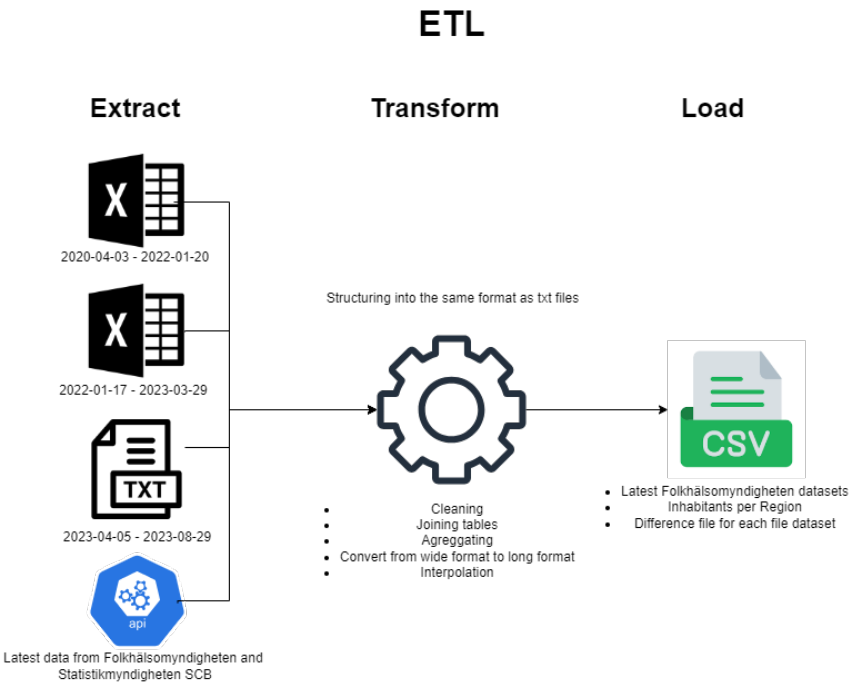

Figure 1: ETL for the Folkhälsomyndigheten data

ETL, stands for Extract, Transform, and Load, is a process in data warehouse involving the extraction of data from various sources, transforming it into a suitable format, and then loading it into a data warehouse (Elmasri and Navathe 2015).

Extraction Phase

Majority of the data has been collected by the researcher providing this project, covering the period from April 3rd, 2020, to September 30th, 2022, in the form of Excel files. From April 5th, 2023, to August 29th, 2023, the data format transitioned to text files with additional details. Latest data can be collected through the Folkhälsomyndigheten API. The data for number of inhabitants per region was collected from SCB, which can be used to calculate the value per 100 000 inhabitants. Therefore, the data for this research project was extracted from Excel files, CSV files, and through a public API from Folkhälsomyndigheten and SCB.

Transformation Phase

During transformation, the data underwent several processes in order to match the format of the latest text files:

- **Cleaning**: Data from different datasets or sheets were cleaned. In cases of ambiguous values (like ".." or "."), these were interpreted as unknown data for years prior to 2020 or as placeholders for future, uncollected data (e.g., projections for 2024 week 50).
- **Joining tables**: Data from different datasets or sheets were combined.
- **Aggregation**: Data was aggregated in order to have same form as text files.
- **Convertion from wide format to long format**: Multiple columns were melted into single columns with various categories or region.
- **Interpolation**: Example can be seen in Table 3, in case of bcov19Kom where data was represented as "<15" in Excel sheets or as ".." in text files, numerical interpolation was used. For instance, if a row showed an increase of 9 cases, implying the previous count was 10, an average was taken to represent this numerically. The interpolation method ensured that the interpolated values can only be whole numbers. This method adjusts the data by handling accumulating fractional components, rounding

them to whole numbers where necessary to maintain the total case count. For interpolation of values things like Kalman Filter smoothing could have been used. Although, the author believes that it would require adapt it to each region and may seem like overengineering in order to get small improvement even though the correct results are unknown. Other interpolation method can be suitable depending on the region, some had the "tot_antal_fall" less than 15 for 30 weeks. Meanwhile, there are regions that can probably be better described by other polynomial equations.

Table 3: Example for Interpolation

Year	Weeknumber	Kommun	tot_antal_fall	nya_fall_vecka
2020	10	Ale	0	0
2020	11	Ale	<15	<15
2020	12	Ale	<15	<15
2020	13	Ale	<15	<15
2020	14	Ale	19	9

Loading Phase

The latest file is stored, which can be seen as the newest data and the most correct data with comparison to the older files of the same type. There would have been multiple of duplicate data which would take space and waste storage. Therefore, there is also a transformed_data which has all the backcorrections of data with values before, after the change and the date of the change. Running the python script within the filder, will automatically store the latest file and update the transformed_data. In addition, there is a municipalities csv file that has the 2022 inhabitant per region numbers. As the latest one for 2023 have not been released.

3.3 Data Mining

The cleaned data was then used for data mining in order to answer the research question. The clustering method Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was applied on the "Aggregated count" and "Aggregated absolute difference" for the municipalities that updated the covid values for number of cases, number of intensive care and number of deaths. As some larger cities may have usually more cases due to the number of inhabitants, normalization was done to get the municipalities on the same scale by dividing the cases by number of inhabitants. As seen in Figure 2 and Figure 5, there are municipalities that are further away from the rest of the data and can be seen as outliers.

There are multiple clustering methods, where the use case depends on the data in hand and the goal of the clustering (Lakshmi and Sahana 2018). For this case, the regions similarity is of interest where the dataset consist of smaller number of sample with outliers. According to Tomar(Tomar 2013), both hierarchical clustering and K-means clustering struggles with outliers.

DBSCAN which is a density based clustering method presented in 1996 (Ester et al. 1996) can be a suitable method. Although, according to Lakshmi and Sahana (Lakshmi and Sahana 2018) DBSCAN may have some weaknesses, that it is sensitive to the parameters. With comparison to other density based clustering method like OPTICS (Ordering Points to Identify Clustering Structure), which is an extension of DBSCAN and results in cluster ordering (Lakshmi and Sahana 2018). DBSCAN simplicity seems to be a better method for this type of data. Although, it needs to be ensured that the parameters are correctly chosen. After, testing different parameters the author decided to set the parameters eps = 500, minPts = 5 for Figure 2 and eps = 0.001, minPts = 5 for Figure 5 as it seems to give reasonable clusters of the regions.

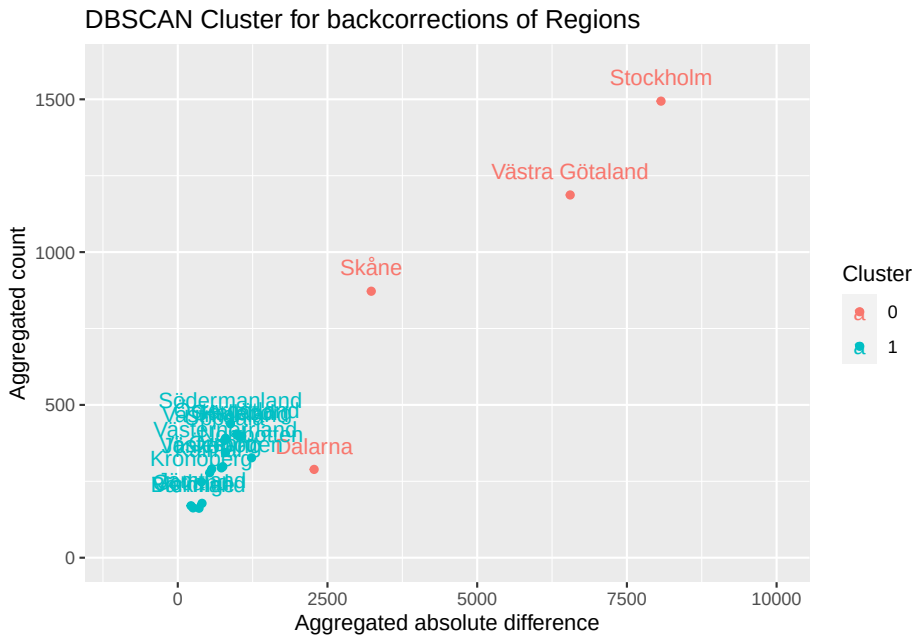

Figure 2: DBSCAN of regions in Sweden

The Y-axis corresponds to aggregated count of backcorrections, which can be defined as the total number of backcorrection instances for each region. Then the X-axis corresponds to the total sum of the all absolute differences between for each backcorrection. This means for every instance where a backcorrection was made, the absolute differences are summed to give a total value.

In Figure 2, the regions were divided into two clusters. Cluster 0 seemed to be the less efficient with comparison to the cluster 1. Regions in Cluster 0, changed the values more frequently and with larger magnitude. Meanwhile, regions in Cluster 1 have usually smaller values and are closer to each other.

Regions such as Stockholm and Gothenburg might more frequently adjust their data due to larger populations. A potential solution is to normalize the data by dividing the values by the number of inhabitants in each region. However, before proceeding with normalization, it's crucial to establish comparability between the number of inhabitants and the variables on the x and y axes. To confirm comparability, the corresponding x and y values against the inhabitants' data are plotted in a scatter plot and linear regression is performed to analyze the correlation.

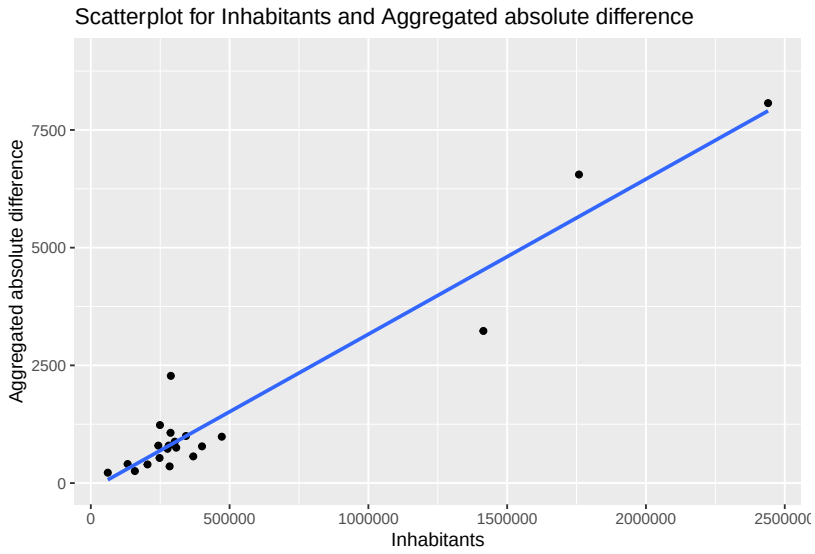

Figure 3: Scatterplot for Inhabitants and Aggregated absolute difference

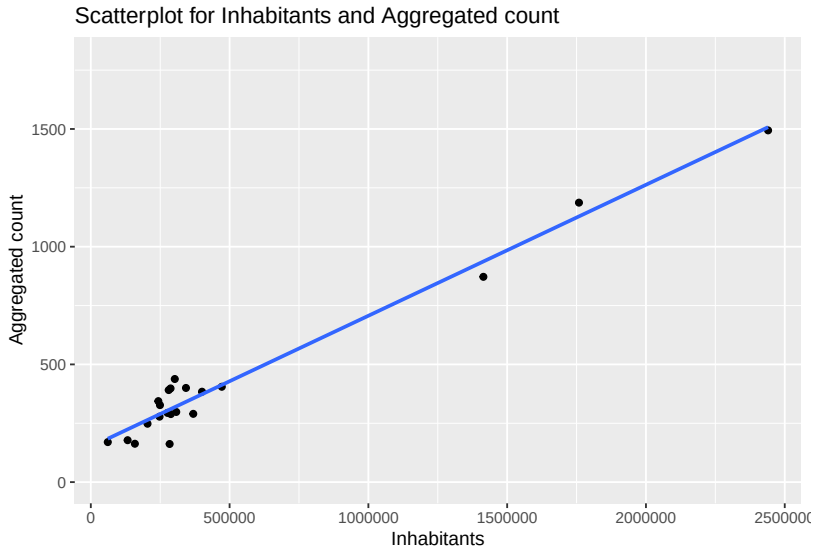

Figure 4: Scatterplot for Inhabitants and Aggregated count

In Figures 3 and 4, the blue line corresponds to linear regression line. This indicates a positive correlation or positive trend: as the number of inhabitants rises, so does the corresponding backcorrection value. This suggests that normalizing the data by the number of inhabitants is a reasonable solution. In this case, it will result in values per capita.

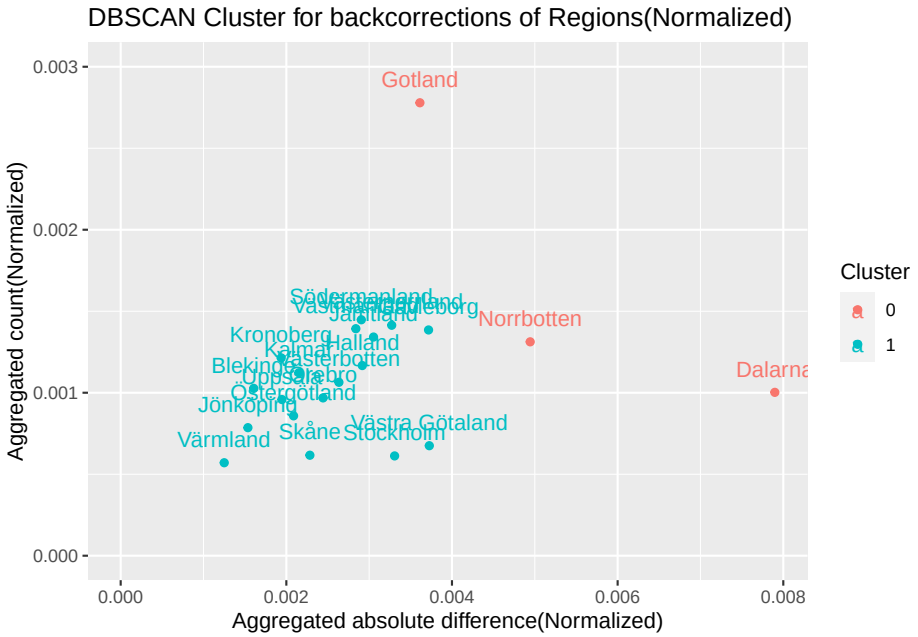

Figure 5: DBSCAN of regions in Sweden (Normalized)

In Figure 5, both axis y and x are same as in Figure 2, but the values are now normalized. After the normalization, the clusters got closer to each other. Regions like Stockholm, Västra Götaland and Skåne that usually corresponds to the largest regions, have lower backcorrection count after normalization and were classified as cluster 1. Instead, Gotland and Norrbotten changed to cluster 0. Regions like Dalarna has been classified as cluster 0 both of the figures.

Important to mention, that it can be hard to fully see the municipalities and the points. It would be required to visualize it in plotly, which would require html or recreating the code in R. Altough this plot is able to show the municipalities that belong to the other cluster and are further away. Presenting the clusters in a table could also have been a possible solution, but it may not give the same typ of understanding as a visual solution.

To answer the research question “How similar are different municipalities in reporting the COVID-19 cases in Sweden?”.

There seem to be similar efficiency between most of the regions in Sweden. As majority of the municipalities in Sweden can belong to one cluster. There are some clusters that seem to be further away from the rest. It would be needed to further analyze the magnitude of the issue and reason for the backcorrection of the data. If the issue is due to human error as discussed in “Inaccurate recording of routinely collected data items influences identification of COVID-19 patients” (Klappe et al. 2022). Municipalities that belong to Cluster 0 may learn from some municipalities from Cluster 1 like Värmland that seem to have been very efficient with reporting the COVID-19 cases.’

Important to mention

Generative AI in form of Github Copilot and ChatGPT was used in this report. The author followed policies stated in course Text Mining (732A81).

”

- Student produces all content, AI only provides inspiration & ideas.
- Student produces all written content, AI helps produce code.
- Student produces all written content, AI helps revise & reformulate.

”

References

Elmasri, Ramez, and Shamkant B Navathe. 2015. *Fundamentals of Database Systems*. 7th ed. Upper Saddle River, NJ: Pearson.

Ester, Martin, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise." In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 226–31. KDD'96. Portland, Oregon: AAAI Press.

"Folkhälsomyndigheten." n.d. https://www.folkhalsomyndigheten.se/the-public-health-agency-of-sweden/.

Klappe, Eva S., Ronald Cornet, Dave A. Dongelmans, and Nicolette F. de Keizer. 2022. "Inaccurate Recording of Routinely Collected Data Items Influences Identification of COVID-19 Patients." *International Journal of Medical Informatics* 165: 104808. https://doi.org/https://doi.org/10.1016/j.ijmedinf.2022.104808.

Lakshmi, Miranda, and Josephine Sahana. 2018. "Review on Density Based Clustering Algorithms for Big Data." *International Journal of Data Mining Techniques and Applications* 7: 13–20. https://doi.org/10.20894/ijdmta.102.007.001.003.

Tomar, Divya. 2013. "A Survey on Data Mining Approaches for Healthcare." *International Journal of Bio - Science and Bio - Technology* 5 (October): 241–66. https://doi.org/10.14257/ijbsbt.2013.5.5.25.