COMP4170: COVID-19 Data Mining

Linpu Zhang
University of Manitoba
Faculty of Science
Winnipeg, Canada
zhangl53@myumanitoba.ca

***Abstract*—During the peak of COVID-19 pandemic we saw large jumps in infections, we delve into the link between the COVID-19 related infection rates and population density in Canada. Our focus includes illustrating the relationship between population density and infection rates in health regions, as well as exploring the connection between infection rates and socio-economic characteristics. Leveraging a comprehensive dataset, we employ visualization, data mining techniques and statistical analysis, providing valuable insights into the dynamics of pandemics.**

Keywords—Data mining, COVID-19, hospitalization, infection, infection rate, population, population density, health region.

I. Introduction And Related Works

The COVID-19 pandemic emerged in late 2019 and rapidly developed into a global health crisis. Using data mining technologies, we hoped to uncover some of these unexplained patterns, such as the relationship between infection rates and hospitalizations, the connection between population density and infection rates, as well as the link between infection rates and socio-economic characteristics. More importantly, using data mining techniques on the COVID-19 in Canada dataset provides an effective and efficient method for revealing non-trivial patterns, and correlations. Our project focuses on the analysis of a large amount of COVID-19 data from Canada (such as infection rates, hospitalizations, cumulative and daily cases in different provinces and health regions, etc.).

In our study, we specifically explored data at the health region level, examining factors like infection rates, population density, and peer groups categorized by principal characteristics. Health regions, being smaller and more numerous than provinces, offer more localized data. This approach enables a deeper understanding of how local social and economic conditions in specific areas influence the spread of the virus.

Moreover, we place significant emphasis on data visualization in our paper. Given the extensive amount of data generated in the healthcare sector, especially concerning epidemic diseases like COVID-19, employing visual representations is crucial. These visuals simplify complex data, making it more accessible and understandable, which is vital in aiding the fight against the disease.

There has been a specialized tool for visualizing and analyzing COVID-19 data that’s been presented, emphasizing the importance of data visualization in understanding complex epidemiological data and its potential applications in other domains[1].

Various interactive dashboards, such as those by Johns Hopkins University, the New York Times provide real-time insights into the pandemic's growth through various formats like bubble maps, choropleth maps, and line heatmaps. Time-series alignment for comparison, a feature aligns time-series data by specific thresholds to compare how different regions have passed through specific stages of the pandemic[2]. There is study to visualize the global spread of the COVID-19 pandemic over the first 90 days, through the principal component analysis approach of dimensionality reduction, correlation matrices and standardized plots[3].

Application of advanced data analytics, AI, and machine learning techniques to predict outbreaks, optimize resource allocation, or improve diagnostic tools has been presented[4]. There has been analysis of Google Trends data in Iran with LSTM models and RMSE to understand public sentiment regarding lockdown measures, vaccine rollouts, and government responses to the pandemic[5].

Supervised machine learning models has been used for diagnosing COVID-19, with an epidemiological dataset from Mexico. It employs algorithms like logistic regression, decision tree, SVM, naive Bayes, and artificial neural network, with a focus on accuracy, sensitivity, and specificity[6]. The findings indicate these models could significantly aid in COVID-19 diagnosis, offering potential relief to strained healthcare systems.

In Canada, each province and territory is responsible for its own healthcare management. This means that COVID-19 data is collected by various provincial health authorities, leading to a diverse and decentralized dataset. Within each province, such as Manitoba, data is collected by different Regional Health Authorities (like the Winnipeg Regional Health Authority and other RHAs). Each RHA collects data relevant to its specific region, which can include infection rates, hospitalization data, testing results, and vaccination rates.

A spatial data science system for analyzing COVID-19 data emphasizes the importance of understanding the geographical aspects of the disease's spread and its implications. The focus is on spatial data analytics across different geographic locations and the spatial hierarchy offers a more efficient, insightful, and user-friendly way to understand and conduct spatial data analytics at higher granularity[7].

II. Data Collection

A. COVID-19 API

The Open COVID-19 API provides a definitive source for data regarding the COVID-19 pandemic in Canada. There are three groups of datasets[8]:

- health region-level (hr), for case and death data only
- province/territory-level (pt), for all data types
- Canada-level (can), for all data types

This enables statistical analysis on all provinces in Canada for various factors contributing to COVID-19. To facilitate the analysis, A small program has been developed to systematically mine a large volume of data from the API, providing daily updates in variables for each day. Variables such as deaths, infections, vaccinations, and more have been mined every day since the start of the pandemic with this API, providing a continuous and latest stream of data for our comprehensive study.

B. Government Statistics

We utilized publicly available Government of Canada statistics “Population and dwelling counts: Canada, provinces and territories” and “Health Regions: Boundaries and Correspondence with Census Geography”. These tables present the population and number of dwellings, land area, and population density for Canada, provinces and health regions in 2021 and 2022. From this table, we extract the data that will help us in our studies, specifically focusing on each province’s, health region’s population and population density. The data provides an important background for our analysis, and it comes from official government sources, representing authority and accuracy. This is an invaluable resource for our study.

III. Data Presentation

The data extracted from the Open COVID-19 API primarily consists of whole numbers. To make meaningful comparisons across different variables, we had to normalize the data. This involved a meticulous normalization process where we carefully selected the start and end dates of the data. Specifically, we utilize normalization techniques to account for population differences among provinces and health regions. The goal is to construct an unbiased level playing field for comparisons of different variables so that we can more accurately and equitably discover patterns and correlations while ensuring that our analyses are not biased by the changes in population size within the COVID-19 dataset. This is the most fundamental indispensable step in building our studies.

Due to a change in the data collection source of the database in September 2023, we excluded data post that date to maintain consistency. Our study concentrates on the timeframe when the virus was active and posed a serious health crisis. As the epidemic diminished and ceased to be a public health concern after mid-2023, it was deemed appropriate to focus our analysis on the period from 2020 to mid-2023.

IV. Results

A. Overall Trend of COVID-19 in Canada

Fig.1 is a representation of the daily new cases of COVID-19 in Canada over a certain time period. The graph exhibits several waves of infection, which is characteristic of the COVID-19 pandemic. These waves correspond to periods where new daily cases rise sharply and then fall.

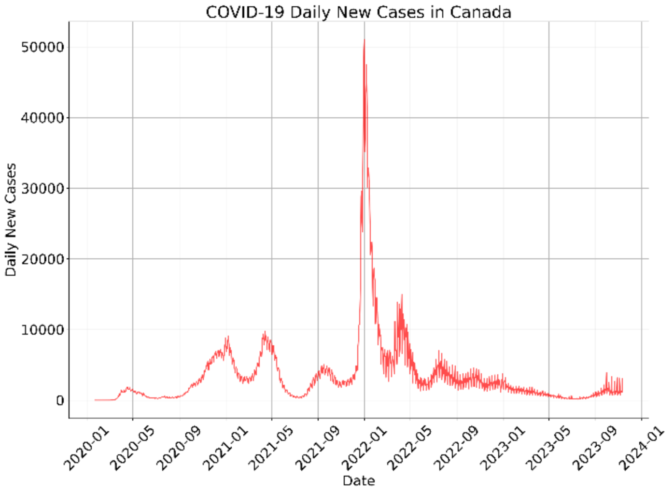


Fig. 1. COVID-19 Daily New Cases in Canada.

The first few waves show a gradual increase in amplitude, with each subsequent peak being higher than the last. This suggests that the virus was spreading more extensively or testing was becoming more widely available and capturing more cases.

The most significant peak in Jan 2022 appears to be an outlier, with a very sharp spike that far exceeds any other point on the graph. This was due to the emergence of a new and more infectious variant Omicron, a super-spreader event, or it could be a data anomaly such as a backlog of cases being reported in a single day due to delayed testing results or reporting.

After the largest peak, the number of daily new cases drops significantly but remains volatile, with several smaller spikes, indicating ongoing transmission and outbreaks.

The "noisy" or jagged appearance of the line, especially in the latter part of the graph, suggests day-to-day variability in reported case numbers. This could be influenced by factors such as testing availability, reporting delays over weekends and holidays, or fluctuations in transmission.

Fig.2 is a line chart that shows the trend of COVID-19 cumulative cases in Canada over time. The x-axis represents the date, and the y-axis represents the number of cumulative cases. The data covers a period from early in the pandemic (January 2020) to September 2023.

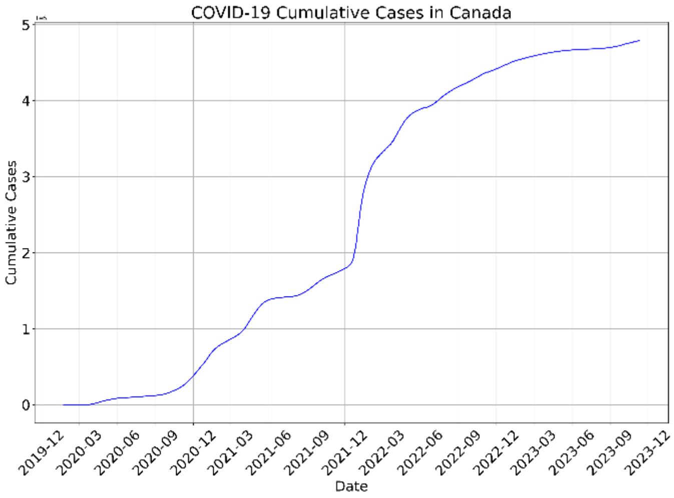


Fig. 2. COVID-19 Cumulative Cases in Canada

The line starts at the lower left, indicating the initial cases of COVID-19 in Canada, and then rises sharply as more cases are reported. This is consistent with the spread of the virus and the reporting of new cases. Curve's slope gives us an indication of the rate of new cases. The Steepest slope happened in Jan 2022, which is consistent with the daily data.

Towards the end of the graph, there is a noticeable leveling off of the curve, suggesting a slow-down in the rate of new cases being added to the cumulative total. This could be due to a variety of factors, including a high proportion of the population gaining immunity (through vaccination or previous infection), effective control measures, or perhaps a reduced spread of the virus.

B. Overall Trend of Cases across Province

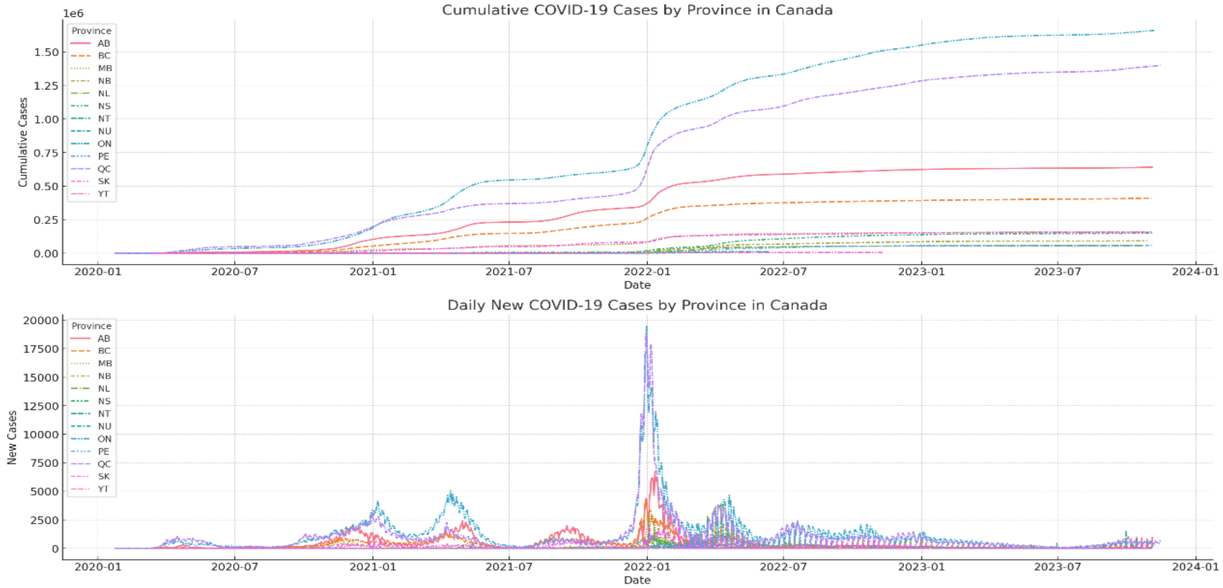


Fig. 3. COVID-19 Cumulative and Daily New Cases by Province

The two graphs in Fig.3 are cumulative and daily wew COVID-19 cases by province in Canada. They clearly show that Ontario and Quebec, being the provinces with the highest populations, naturally report the greatest number of COVID-19 cases.

Notably, these provinces also experienced the most pronounced surge of infections in January 2022, characterized by a steep increase in case numbers.

C. Cases and Hospitalizations

Finding a sequential mathematical relationship between daily new COVID-19 cases and hospitalizations is a complex task, as the relationship can be influenced by numerous factors. However, a common approach to explore such relationships is through correlation and time series analysis.

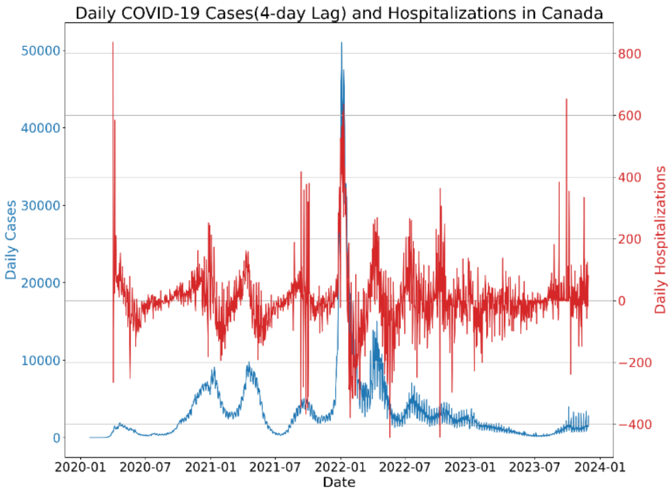


Fig. 4. COVID-19 Daily Cases with 4-day Lag and Hospitalizations in Canada

The Pearson correlation analysis involves calculating the correlation coefficient between daily cases and hospitalizations. A strong correlation might indicate a direct relationship[9].

The Pearson correlation coefficient between daily new COVID-19 cases and daily hospitalizations in Canada is approximately 0.37. This indicates a moderate positive correlation. In other words, there is some degree of linear relationship between the two, but it's not very strong[9].

Often, hospitalizations lag behind cases because there's a delay between when people get infected, when they develop symptoms severe enough to require hospitalization, and when these hospitalizations are reported. We can analyze this by introducing a lag in the cases data and then calculating the correlation.

The typical time lag could range from a few days to a couple of weeks. The strongest correlation between daily new COVID-19 cases and hospitalizations in Canada occurs with a 4-day lag as shown is Fig.4, and the Pearson correlation coefficient at this lag is approximately 0.41. This suggests a slightly stronger relationship than the immediate day-to-day correlation, indicating that hospitalizations tend to follow the trend in cases with a delay of about 4 days.

It's important to note that correlation does not imply causation and that other factors might influence these trends. Additionally, more sophisticated time series analyses could uncover more complex relationships, but this gives a general idea of how the two metrics are related.

D. *Infection Rates and Population Density across Health Regions*

In our study, we use health regions instead of provinces to analyze covid data. Health regions are administrative areas defined by provincial or territorial governments in Canada. They are used for organizing and delivering health services, public health reporting, and health resource allocation. They are the smallest area unit we can get for covid infection data. According to the data in 2018, Canada had 100 health regions.

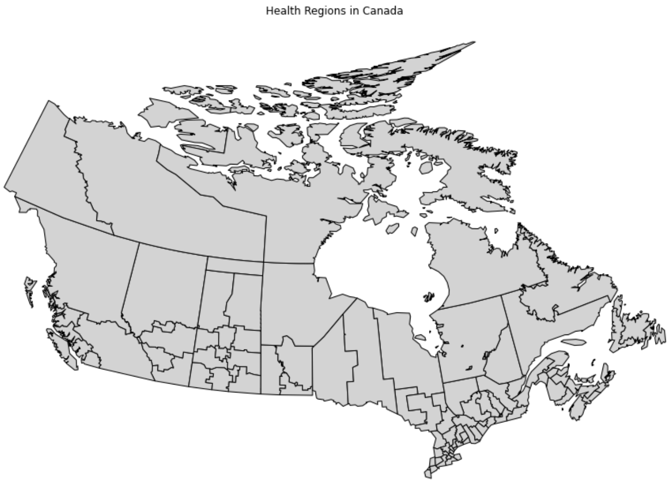
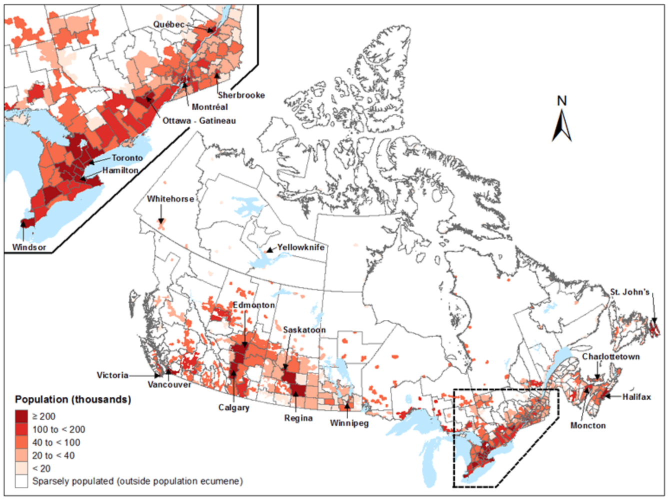


Fig. 5. Health Regions in Canada

In recent years there has been an increasing demand for relevant health information at a 'community' level. As a result, health regions have become an important geographic unit by which health and health-related data are produced. Being smaller and more numerous than provinces, they provide more localized data. This allows for a better understanding of how a disease is spreading in specific areas, which can be crucial for targeted public health interventions.



Source: Statistics Canada, Demography Division

Fig. 6. Population Density in Canada

Fig.6 provides a visual representation of population distribution across Canada, segmented by the density of the population in thousands within various regions.

Major cities such as the Greater Toronto Area, Montreal, and Vancouver are the largest population centers. The population is heavily concentrated along the Canada-U.S. border, with significant clustering in the provinces of Ontario and Quebec, as well as along the Pacific coast in British Columbia. The northern regions of Canada, including much of the territories (Yukon, Northwest Territories, and Nunavut), are shown as sparsely populated.

While there are population centers in the Atlantic provinces such as Halifax and St. John’s, the overall population density in these regions is lower than in central Canada.

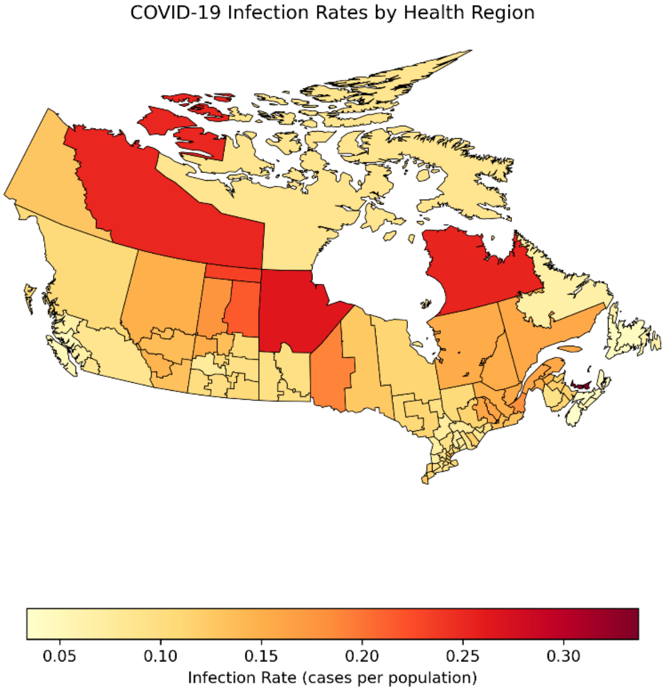

COVID-19 Infection Rates by Health Region
0.05
0.10
0.15
0.20
0.25
0.30
Infection Rate (cases per population)

Fig. 7. Infection Rates by Health Regions in Canada

It visually distinguishes between urbanized areas and rural or wilderness areas, with a clear majority of Canadians living in urban areas.

Fig.7 provides a visual representation of infection rates across health regions across Canada. There are clear variations in infection rates between different regions. Some regions show significantly higher rates, indicating more widespread infection relative to their population sizes. These differences could be influenced by a variety of factors including population density, local public health policies, and access to healthcare services.

The infection rate will be calculated as follows:

$$\text{Infection Rate} = \frac{Total\ Number\ of\ Cases}{Total\ Population} \times 100000 \qquad (1)$$

E. Comparison between Health Regions within Provinces

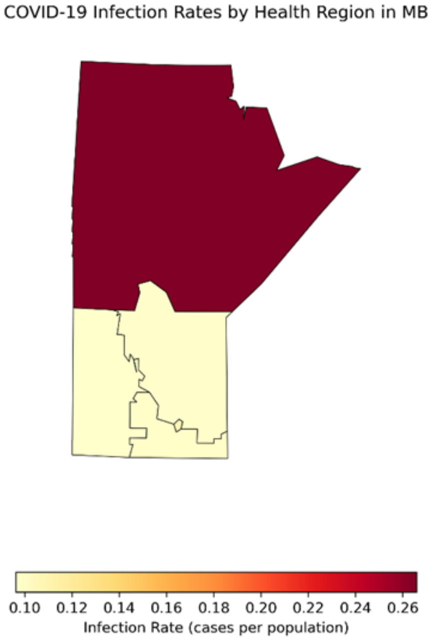

COVID-19 Infection Rates by Health Region in MB
0.10
0.12
0.14
0.16
0.18
0.20
0.22
0.24
0.26
Infection Rate (cases per population)

Fig. 8. Infection Rates by Health Region in Manitoba

It's somewhat counterintuitive that many rural areas have higher infection rates than urban areas, as illustrated by the map. Take Manitoba as an example: the Northern Health Region exhibits the highest infection rate, significantly surpassing that of Winnipeg.

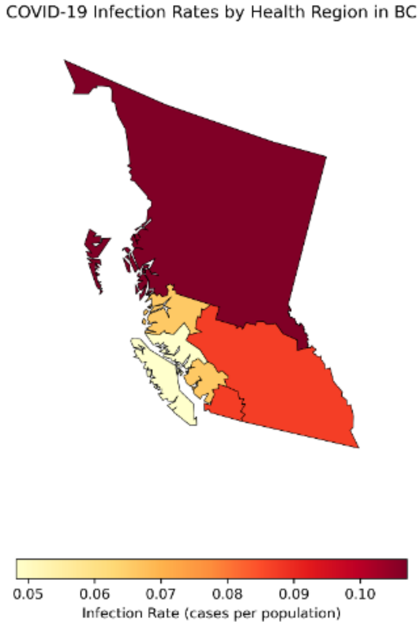

COVID-19 Infection Rates by Health Region in BC
0.05
0.06
0.07
0.08
0.09
0.10
Infection Rate (cases per population)

Fig. 9. Infection Rates by Health Region in British Columbia

This pattern also applies to BC and ON.

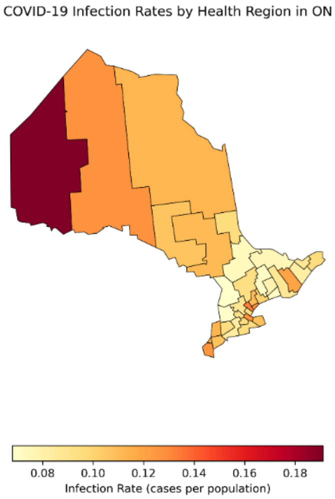

COVID-19 Infection Rates by Health Region in ON
0.08
0.10
0.12
0.14
0.16
0.18
Infection Rate (cases per population)

Fig. 10. Infection Rates by Health Region in Ontario

The pattern becomes clearer when examining the graphs that compare the infection rates and population densities of health regions across various provinces.

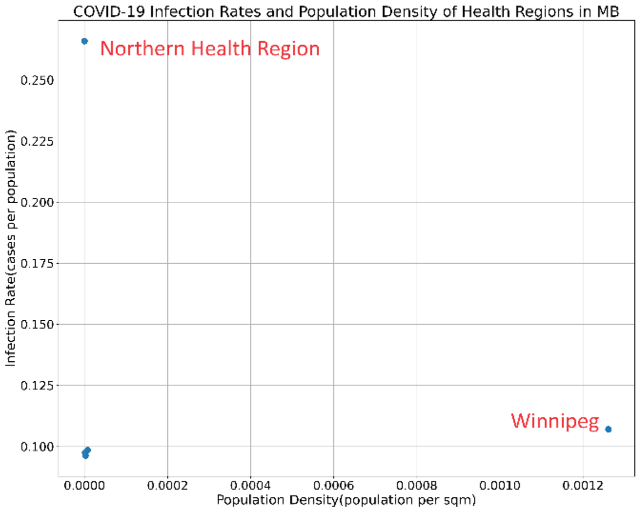


Fig. 11. Infection Rates and Population Density by Health Region in Manitoba

In Manitoba, the Northern Health Region is an outlier according to infection rate, significantly surpassing that of Winnipeg. Winnipeg is the largest city in MB and the population density is much greater than other health regions, but it does not automatically translate to high infection rates. Actually, its infection rate is very close to other health regions with more sparse population.

In Ontario, mirroring the situation in Manitoba, the City of Toronto Health Region does not display the highest infection rate. Instead, it is the Northwestern Health Region, predominantly rural, that has recorded the highest rate. Similarly, the Thunder Bay Health Region, another rural area, also shows a relatively high rate of infection.

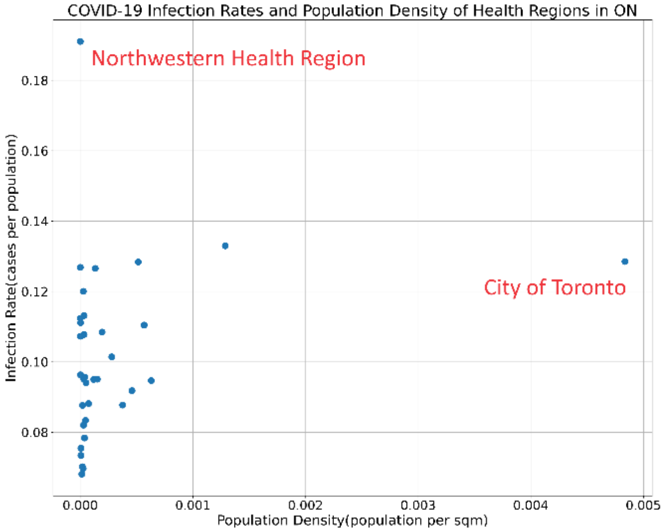


Fig. 12. Infection Rates and Population Density by Health Region in Ontario

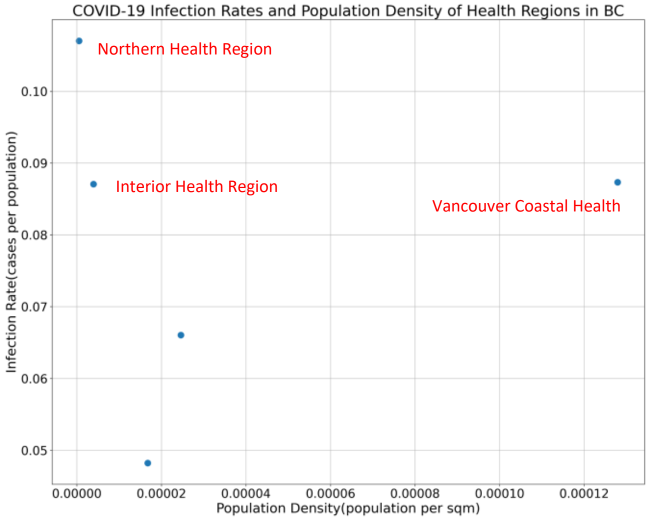


Fig. 13. Infection Rates and Population Density by Health Region in BC

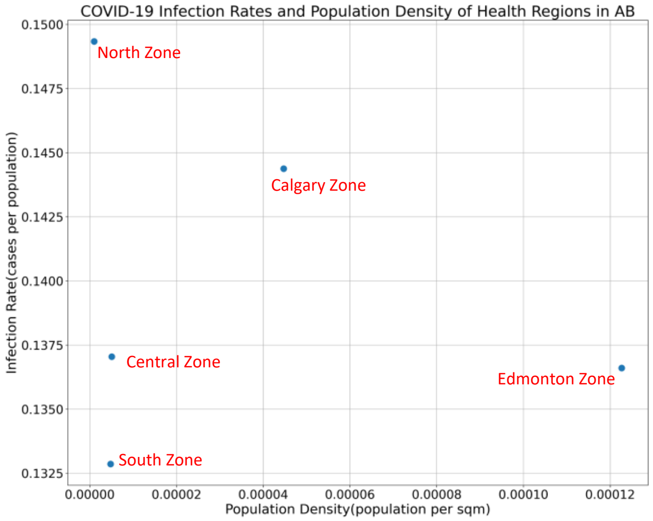


Fig. 14. Infection Rates and Population Density by Health Region in Alberta

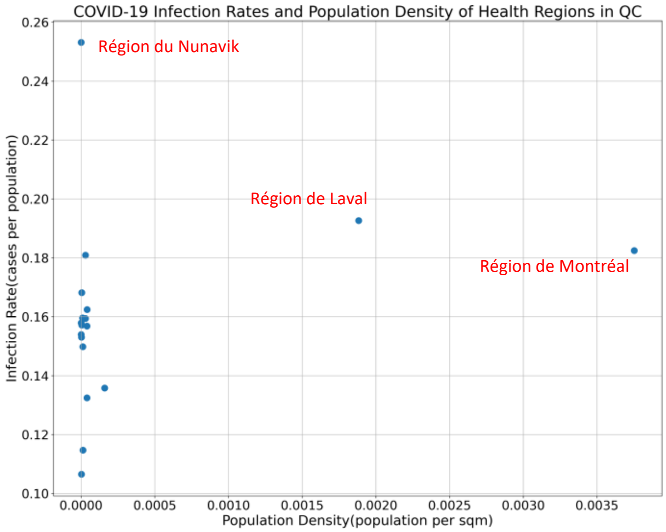


Fig. 15. Infection Rates and Population Density by Health Region in Quebec

In nearly every Canadian province, a frequent pattern emerges. Major population centers such as Vancouver, Montréal, Toronto, and Edmonton have maintained moderate or even relatively low infection rates compared to less populated areas. On the other hand, regions like the Northern Health Region, Région du Nunavik, North Zone, and other remote areas, despite their low population density, have experienced significantly higher infection rates.

Several factors contribute to this phenomenon. First, the majority of Canada's rural population resides in small towns, where the density can actually be relatively high. Second, some rural areas experience colder temperatures year-round, creating conditions more conducive to the spread of viruses. Also, regions with robust healthcare infrastructure might have been more effective in testing and controlling the spread of the virus, potentially resulting in lower infection rates.

This observation clearly indicates that population density alone does not fully explain the variations in infection rates. To better understand the relationship between infection rates and the social and economic conditions in different areas, additional data and factors pertaining to health regions beyond just population density are required.

F. Health Region Peer Groups

In order to effectively compare health regions with similar socio–economic characteristics, health regions have been grouped into 'peer groups'. Statistics Canada used a statistical method to achieve maximum statistical differentiation between health regions. Twenty–four variables were chosen to cover as many of the social and economic determinants of health as possible, using data collected at the health region level mostly from the Census of Canada. Concepts covered include[]:

- Basic demographics (for example, population change and demographic structure)
- Living conditions (for example, socio-economic characteristics, housing, and income inequality)
- Working conditions (for example, labour market conditions)

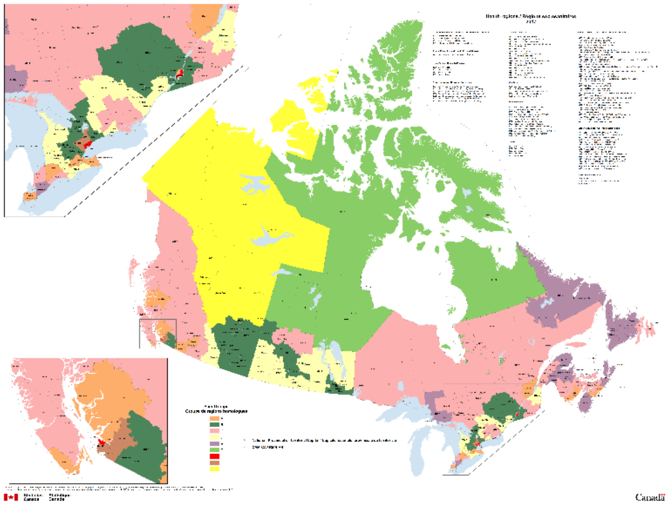


Fig. 16. Health Regions and Peer Groups in Canada[9]

Peer groups based on 2017 health region boundaries and 2011 Census of Population and 2011 National Household Survey data are available. There are currently nine peer groups identified by letters A through I. There have been no changes made to peer group assignments since 2014[10].

TABLE I. TABLE OF PEER GROUPS AND PRINCIPAL CHARACTERISTICS[10]

Peer group	Number of health regions	Percent of Canadian population	Principal characteristics
A	13	15.7	• Population centres with high population density and rural mix from coast to coast • High percentage of visible minority population • Low percentage of Aboriginal population • Average employment rate
B	19	31.2	• Mainly population centres with moderate population density • Average percentage of visible minority population • High employment rate
C	31	14.5	• Population centres and rural mix from coast to coast • Average percentage of visible minority population • High percentage of Aboriginal population
D	18	6.5	• Mainly rural regions in Ontario and the Prairies • Low percentage of visible minority population • Average percentage of Aboriginal population
E	12	3.2	• Mainly rural Eastern regions • Low percentage of visible minority population • Low employment rate
F	5	0.5	• Northern and remote regions • Very low percentage of visible minority population • Very high Aboriginal population
G	3	15.3	• Largest population centres with an average population density of 4211 people per square kilometre • High percentage of visible minority population • Very low Aboriginal population
H	5	11.5	• Mainly population centres in Ontario and British Columbia with high Population density • Very high percentage of visible minority population

			• Low Aboriginal population
I	4	1.7	• Mainly rural and remote regions in the Western provinces and the Territories • Average percentage of visible minority population • High percentage of Aboriginal population • High employment rate

Fig.17 is the graph representations of infection rates by health region and the peer group it belongs to. Each point on the graph represents a health region. The x-axis shows the infection rate for each health region. The y-axis corresponds to the province that the hr is in. Different colors and markers indicate the peer groups.

This visualization makes it clearer to compare the infection rates across different health regions in provinces and observe any patterns or clusters within peer groups.

It's obvious that population centers (e.g. Peer Group A, B and C) usually with high employment rates and urban areas has moderate or low infection rates. On the other hand, the Peer Group F and I have high or relatively high infection rates within their provinces.

Recall the principal characteristics of Peer Group F and I. Peer Group F:

- Northern and remote regions
- Very low percentage of visible minority population
- Very high Aboriginal population

Peer Group I:

- Mainly rural and remote regions in the Western provinces and the Territories
- Average percentage of visible minority population
- High percentage of Aboriginal population
- High unemployment rate

G. *Conclusion*

This observation raises some important points about the dynamics of infectious disease spread in different settings.

Some high population density regions don't have very high infection rates, which suggests that high population density does not automatically translate to high infection rates. It could be indicative of effective public health measures in densely populated areas, such as more rigorous testing, tracing, and isolation protocols. It might also reflect better healthcare infrastructure and resources in more densely populated areas, which can help manage and control outbreaks more effectively.

Contrarily, Peer Groups F and I, which have significant Aboriginal populations and being in northern, rural, or remote regions, have higher COVID-19 infection rates. This pattern could be influenced by several factors:

- Healthcare Resources: In regions with low population density, healthcare resources might be more limited, and

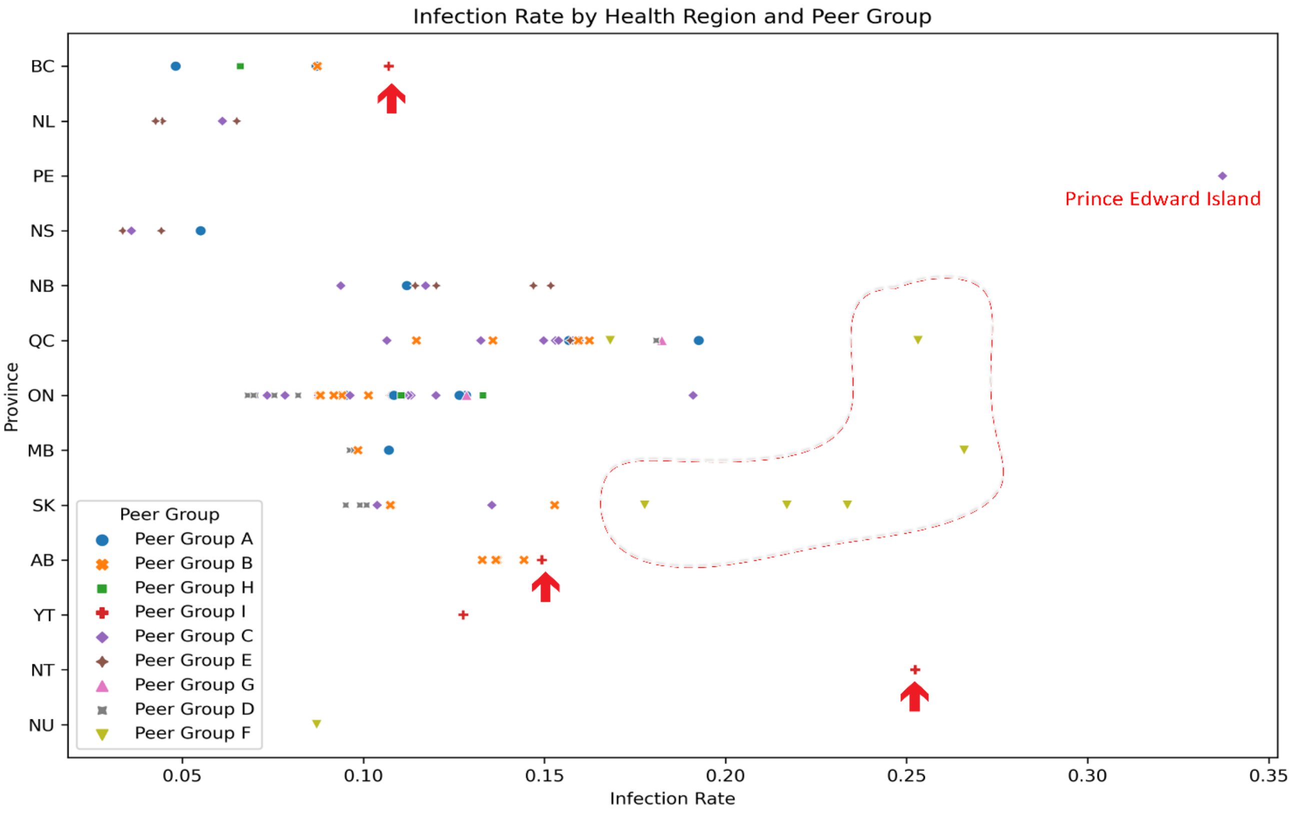


Fig. 17. Infection Rates by Health Region and Peer Group

access to medical care could be challenging, potentially leading to higher infection rates if an outbreak occurs.

- Living Conditions: Overcrowding in housing is more common in some Aboriginal communities, which can facilitate the spread of respiratory infections like COVID-19.
- Comorbidities: High rates of chronic diseases, which are more prevalent in some Aboriginal populations, can increase the risk of severe COVID-19 infections.
- Public Health Infrastructure: There may be limitations in public health infrastructure and resources needed to implement widespread testing, contact tracing, and isolation measures.
- Employment: High employment rates could indicate that a larger percentage of the population is engaged in essential work, where there is a higher risk of exposure to the virus due to a lack of options to work from home.
- Mobility and Travel: In some remote communities, residents must travel to access services, including healthcare, which can increase the risk of exposure to and spread of the virus.
- Cultural Practices: Social and cultural practices that involve community gatherings could also contribute to higher transmission rates if these continue during a pandemic.

- Information and Education: Disparities in education and access to information can affect how communities respond to public health advisories, including understanding the importance of and adhering to guidelines such as physical distancing and mask-wearing.
- Resource Allocation: The allocation of resources for public health measures, including vaccination efforts, may not be equitable, potentially leaving these communities more vulnerable.
- Socioeconomic Factors: Economic disparities can impact the ability of individuals to take time off work when sick, self-isolate, or access healthcare, which can contribute to higher transmission rates.

These patterns underline the complexity of infectious disease dynamics. Factors like public health policy, community compliance with health guidelines, healthcare infrastructure, and even socio-economic factors play significant roles in determining how an infectious disease spreads within a community. This complexity is why tailored approaches are often necessary for different regions, considering their unique circumstances and challenges.

References

[1] C. K. Leung, Y. Chen, C. S. H. Hoi, S. Shang, Y. Wen and A. Cuzzocrea, "Big Data Visualization and Visual Analytics of COVID-19 Data," 2020 24th International Conference Information Visualisation (IV), Melbourne, Australia, 2020, pp. 415-420, doi: 10.1109/IV51561.2020.00073.

[2] J. L. D. Comba, "Data Visualization for the Understanding of COVID-19," in Computing in Science & Engineering, vol. 22, no. 6, pp. 81-86, 1 Nov.-Dec. 2020, doi: 10.1109/MCSE.2020.3019834.

[3] Teh, Jane KL, et al. "Multivariate visualization of the global COVID-19 pandemic: A comparison of 161 countries." PLoS One 16.5 (2021): e0252273.

[4] Roy, Anit N., et al. "Prediction and spread visualization of COVID-19 pandemic using machine learning." (2020).

[5] Ayyoubzadeh, Seyed Mohammad, et al. "Predicting COVID-19 incidence through analysis of google trends data in Iran: data mining and deep learning pilot study." JMIR public health and surveillance 6.2 (2020): e18828.

[6] Muhammad, L. J., et al. "Supervised machine learning models for prediction of COVID-19 infection using epidemiology dataset." SN computer science 2 (2021): 1-13.

[7] S. Shang, C. K. Leung, Y. Chen and A. G. M. Pazdor, "Spatial Data Science of COVID-19 Data," 2020 IEEE 22nd International Conference on High Performance Computing and Communications; IEEE 18th International Conference on Smart City; IEEE 6th International Conference on Data Science and Systems (HPCC/SmartCity/DSS), Yanuca Island, Cuvu, Fiji, 2020, pp. 1370-1375, doi: 10.1109/HPCC-SmartCity-DSS50907.2020.00177.

[8] Berry, I., O'Neill, M., Sturrock, S. L., Wright, J. E., Acharya, K., Brankston, G., Harish, V., Kornas, K., Maani, N., Naganathan, T., Obress, L., Rossi, T., Simmons, A. E., Van Camp, M., Xie, X., Tuite, A. R., Greer, A. L., Fisman, D. N., & Soucy, J.-P. R. (2021). A sub-national real-time epidemiological and vaccination database for the COVID-19 pandemic in Canada. Scientific Data, 8(1). doi: https://doi.org/10.1038/s41597-021-00955-2

[9] Sedgwick, Philip. "Pearson's correlation coefficient." Bmj 345 (2012). J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[10] "Health Region Peer Groups Working Paper." Health Reports (Statistics Canada), 2018, Statistics Canada – Catalogue no. 82-622-X