Open Data Analysis for the City of Surrey

Ziqing Yuan

The University of British Columbia

Dr. Rachel Pottinger

March 20th, 2023

Introduction

Open data sources are increasingly recognized as valuable resources for research and decision-making. Governments often choose to make some of their data open to promote transparency and accountability in public spending, as well as to attract historical data for future planning and construction. To ensure the usefulness of these open data sources, it is crucial that they are well-organized, aligned, and clearly annotated. This personal research project has two main motivations. First, the study aims to examine how open data can be improved to be more aligned and user-friendly. Second, it aims to provide insights for future planning by analyzing historical data. Overall, it offers an opportunity for the author to explore the differences between conducting research and completing a course project.

Data Sources

For this study, we used data sources directly downloaded from the open data catalog maintained by the City of Surrey. The datasets selected for analysis were restaurants, transit, and places of interest, chosen for their data quality and inner correlation.

While the open data platform is free and accessible to the public, not all datasets are relevant to the research purpose, and some are outdated. However, the three selected datasets are continually updated on a monthly basis. Although the transit data was last updated six months ago, it is still considered reliable as transit stop locations tend to be stable from a higher-level perspective.

These three datasets are highly correlated in terms of their locations. Decision-making processes such as setting new transit stops or public facility locations by the government, and small businesses choosing restaurant locations in densely populated areas, are essential to the formation of clustered communities or neighborhoods. Analyzing these datasets together can not only provide a comprehensive overview of historical planning data and existing results but also highlight existing neighborhood centers and potential areas for future development. This may also offer some valuable insights for government officials and members of the public involved in city planning or accommodation planning.

It is worth noting that the datasets are offered in several file formats, but there is no guarantee that all formats align with each other on the website. However, this is not the focus of this research. All of the downloaded data is in the format of the first link found on each respective webpage.

Restaurants

The Restaurants dataset contains restaurant information extracted from Fraser Health restaurant inspection data. The City of Surrey GIS section added latitude and longitude values to the data, and it was downloaded in CSV format.

Transit

The Transit dataset contains public transit data, including skytrain guideways, transit routes, and transit stops. We downloaded the data in JSON format.

Places of interest

The Places of Interest dataset includes public and private facilities operating or located within the City of Surrey, including some private facilities with recycling facilities or drop-off centers. We downloaded this data in CSV format.

Data Processing

This research was conducted using Python. Each dataset was first imported into the program and processed to abstract instances of Python class. Each property/column was converted to a field of the corresponding class instance. This step also included data cleaning and warehousing, with some of the problems and difficulties we encountered listed separately below. We also provide potential solutions for each of the difficulties we encountered. However, these suggestions are based on a high-level perspective and may not consider every single aspect. After processing, due to time limitations for this study, we decided to discard the Places of Interest dataset for further analysis.

Restaurants

This dataset had a clear structure and clean data, making it a good fit for our study. No confusion or problems arose during processing.

Transit

The transit data source is provided in JSON format, and we have mainly encountered structural problems when processing this data.

For the purpose of this study, we only needed the transit stop data. As described above, this dataset also contains data about SkyTrain gateways and transit routes. The way this dataset arranges the three different types of data is by putting them into an array and having one property nested as a property indicating their type. This way of implementation may be reasonable if people need every piece of the data in this dataset and could simply iterate through this array to access each type of data. However, it is not very easy to maintain, as all of the related information is kept in a single file. While many properties of these three data types have different names and types, we believe it might be better to keep them separately in different files. This also comes with the consideration that there might be some more different types of transit data in the future. Besides, it is not easy to understand, especially when there is no other documentation provided aside from this file. This brings out another potential hazard of this dataset.

As no documentation is provided, some of its property namings are not very direct, and their values are also very confusing. It is definitely a good idea to use a number to represent some highly repeated string data, but this is all based on a well-maintained documentation.

Another crucial but resolvable problem in this dataset is about data alignment. For example, within the transit stop data array, their data for the "LOCATION" property is not aligned. Stops outside Surrey city use this field to record their belonging city, while stops inside Surrey city use this to record the specific location on the road. Our suggestion is that they could be separated into two fields. The specific location may also be essential as they might also be using the name of a bus stop.

Places of interest

The Places of Interest dataset had a well-defined structure, but its data quality was relatively low. It included numerous duplicated rows and unclear classifications for some attributes. A detailed description is provided in Appendix A. Since this dataset is relatively large and may require manual data cleaning for further analysis, we did not proceed with additional processing or analysis on this data

Data Analysis and Visualization

The data analysis in this study primarily focuses on the geographical coordinates of restaurants and transit stops, as the Places of Interest dataset was discarded in the previous step.

We analyzed the correlation between the locations of restaurants and transit stops using both graphical and numerical methods.

For graphical analysis, we chose to use a heat map as it provides a direct visualization of how the two types of locations are correlated with each other.

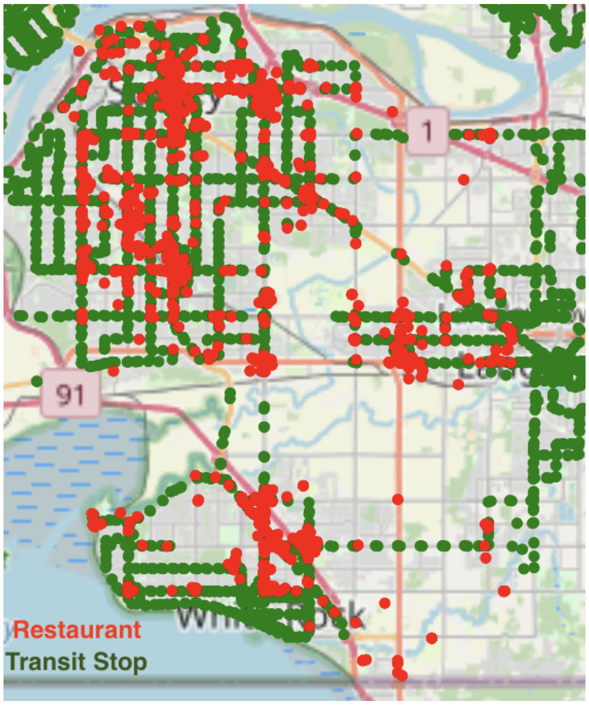

Figure 1. Heat map of Restaurant locations and Transit Stop locations in Surrey

For numerical analysis, we considered two different coordinate systems: the UTM system, originally developed for military use and known for its high accuracy, and the Lat-Lon system, commonly used when referring to coordinates on Earth. We converted all the coordinate data to the UTM system for higher accuracy, and then produced a distribution. Additionally, we calculated the difference between the baricenter and spread to complement our analysis.

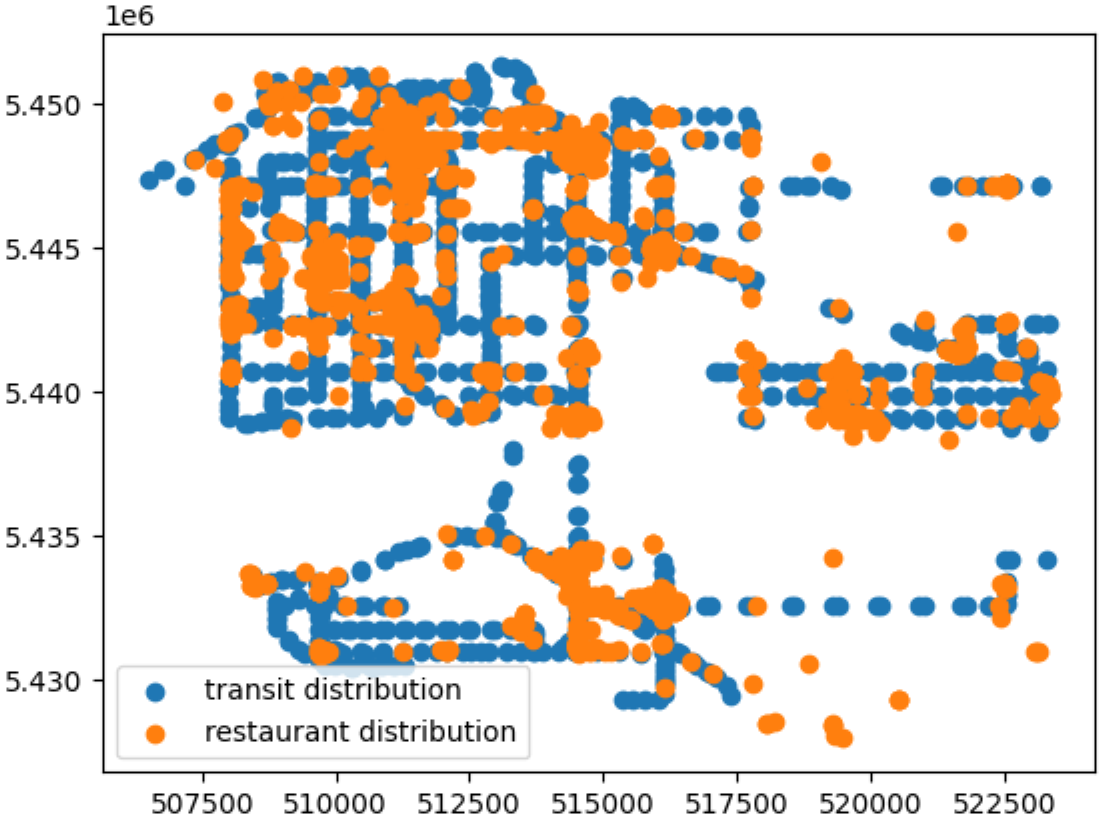

Figure 2. Distribution of Restaurants locations and Transit Stop locations in Surrey

The analysis of the correlation between restaurant and transit stop locations has yielded significant results, confirming the hypothesis of a strong correlation. These findings suggest a crucial interplay between transportation and commercial activities in Surrey city, which could provide some insights for urban planners and small businesses in future planning and development.

Discussion

In conclusion, our research processed and analyzed the restaurant and transit datasets, identifying a strong correlation between these two types of locations. This finding has important implications for data maintenance, urban planning, and transportation infrastructure development in Surrey City. However, our study had some limitations, including the small sample size and exclusion of the places of interest dataset. Future research could explore these limitations in greater depth and consider additional datasets and analytical techniques to gain a more comprehensive understanding of the relationship between urban development and economic prosperity.

References

City of Surrey. (n.d.). Open data. Retrieved from https://data.surrey.ca/

Appendix A. Analysis of Data Processing Challenges

transit.json

The problem I run into when processing	**What causes that**	**Situations where such "problem" may be good**	**Possible solution for future data governance**
It was difficult to access the transit_stops data.	The overall structure was an array with three object elements. The type of information for each element was hidden inside the element and was hard to access if targeting a certain type of data (in this case, transit_stops).	If the research/project needs to process all the data of skytrain_guideways, transit_routes, and transit_stops data, then this data structure would be perfect for iteration and access fields.	Split the data source into separate files based on the information they contain. Doing so could also enable potential future expansions of new data types.
It was difficult to browse the structure of the JSON file.	There was no documentation provided to describe the structure of the JSON file. Some of the namings were confusing ("type" : "Feature") and some of the variable meanings were unclear (e.g., "TYPE" : 3).	N/A	Provide a description or document for the JSON structure information.
The data are not aligned in some fields.	Data for the same property was not consistent. For example, the "LOCATION" field stored the city location for a transit stop that was not in Surrey city (e.g., "BURNABY"), but stored the stop name for a transit stop that was in Surrey city	The dataset also including transit data in surrounding cities could be helpful for providing citizens with intuition for how far they could go by taking public transit from Surrey.	Distinct two different types of data in the same field, like "stop_name" and "city," rather than combining them.

	(e.g., "Northbound 123A St @ 98 Ave").		
There was some inconsistency observed.	Even though the overall structure was an array, the elements' fields were not aligned within that array.	If the research/project needs to process all the data of skytrain_guideways, transit_routes, and transit_stops data, then this data structure would be perfect for iteration and access fields.	Keep the fields consistent among elements.

restaurant.csv

This table had a very clear structure and clean data, and it fit my analysis well.

places_of_interest.csv

The problem I run into when processing	**What causes that problem**	**Possibilities that this "problem" may be good**	**Possible solution for future data governance**
Some places' names were null.	N/A	N/A	Use some third-party APIs or visit that place to find their names.
Some place types in the Type column were highly related, and some places' types were null.	N/A	N/A	Create some abstraction for the type and make those types subtypes. For example, "Tennis" and "Volleyball" should not be directly put in as types but can be subtypes for "Sports." Also, if there is some facility with an uncleared type, it should be

			categorized as “Other” but not null.
There was some degree of inconsistency between Location(address) data and Lon-Lat data.	Some locations were null but latitude and longitude had data.	N/A	When collecting the initial data and facing that situation, resolve it first-hand (by either using Google Maps or making that field not null).

Appendix B. Data Processing Solutions Table

General Problem	Example	How I solved it
Inconsistency in the semantics of data within individual data sources.	*transit.json:* The 'LOCATION' field in the transit stop data contains different types of information depending on the record. For transit stops located within Surrey, the 'LOCATION' field stores the name of the stop, such as 'Northbound 123A St @ 98 Ave'. For transit stops located outside of Surrey, the 'LOCATION' field stores the name of the city where the stop is located, such as 'BURNABY'.	Since all city names in the existing data are in capital letters, I created a separate field called 'CITY' to store the corresponding city names. I transferred all the city names to this field while still retaining the location of stops in Surrey in the original “LOCATION” attribute. The code for this process can be found in the *json_to_stop_data* function within the ./scripts/data_processing.py file in the repository.
Incompatibility in units across disparate data sources.	*transit.json and restaurant.csv:* Thy use different coordinate systems for their locations.	I developed a function to abstract functionalities from a GIS library for converting between the different coordinate systems used in those two sources. The function is available in the *map_system_converter* function located within the ./scripts/data_processing.py file in the repository.