The University of Queensland

Visualisation Project

South East Queensland: Translink™ Data

Ryan Phelan

43935707

Semester 1 2017

COSC3000 Visualisation, Computer Graphics and Data Analysis

Due Date Tuesday April 25th, 4pm

Contents

Introduction

Public transport is a vital part of populated cities and regional towns. Maintaining a constant flow of transportation throughout a city means that more people can commute to and from their places of business and tourists/locals can find cheap methods of travel, this is crucial for the economic growth of a region. Buses and trains that can carry many passengers are both necessary to reduce urban traffic levels and regional carbon emissions.

In South East Queensland, the Translink™ agency of the Department of Transport and Main Roads currently operates to coordinate public bus, train and ferry routes for the general populace. Services operate year-round, in 2010 alone Translink™ carried over 77 million passengers (Hurst, 2011). Trends in transportation are important for urban developers, government road planners, local businesses and of course, the local citizens that rely on it for their daily commute.

Data

For this project, I was able to procure a wide variety of data generated directly by the Translink™ agency. Their main website offers an "Open Data" section and is free to the general public for downloading and using. Notable data files include:

- Stop Information
  - Stop Names
  - Stop ID's
  - **Latitude and Longitude** for each Bus, Train and Ferry Stop
- Stop Times
  - Every instance of a bus arriving/departing a location
- Go-Card© Technology Data
  - List of every Go-Card© retailer (Locations included)
  - Total Trips taken using a Go-Card©
- Travel Data
  - Total number of trips for each week

The data was downloaded from https://translink.com.au/about-translink/open-data. Files were all downloaded as text files and converted to Comma Separated Value (csv) when needed.

Aims

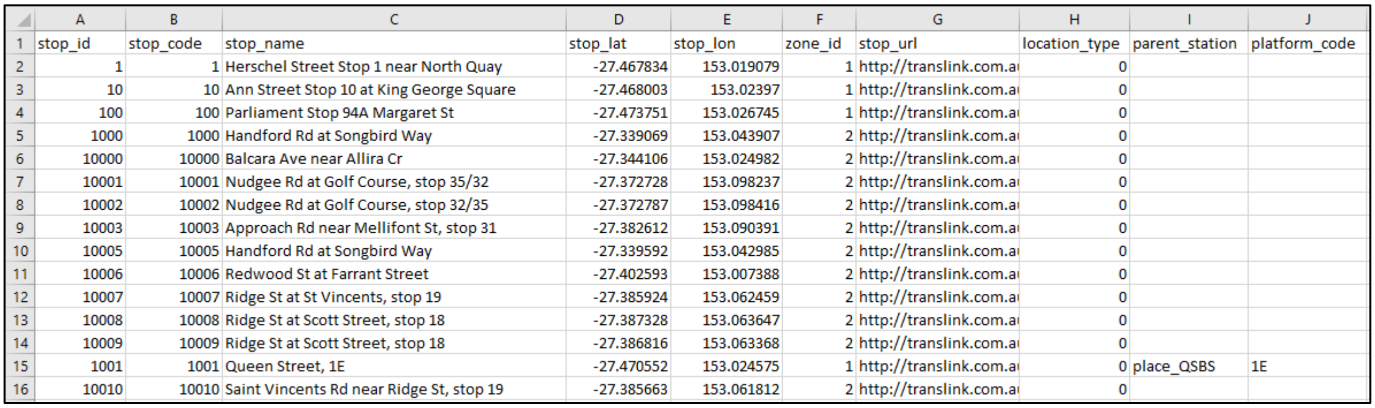
My main goal of the project is to create a variety of useful data visualisations across a number of software mediums that will give insight into both the effectiveness and the extent of Translink™ transport. Also, finding ways that SE Queensland transport might be improved overall. Specific areas for data visualisation include:

- Visualising transport usage with respect to the time of the year
- How much travelling is done using Go-Card© technology
- Where are the easiest areas to buy a Go-Card©
- What time of day is Translink™ most active
- What is the distribution of stops across SE Queensland
- Where is public transport the most active

Methods

Data Source

The files acquired through the Translink™ website were all basic text files, however ultimately their inner formatting was that of a csv file. Data entries each had a variety of attributes including name, id, and location. A snippet from stops.txt is shown below, after it has been imported to Microsoft Excel™ as a csv file.



	A	B	C	D	E	F	G	H	I	J
1	stop_id	stop_code	stop_name	stop_lat	stop_lon	zone_id	stop_url	location_type	parent_station	platform_code
2	1	1	Herschel Street Stop 1 near North Quay	-27.467834	153.019079	1	http://translink.com.a	0		
3	10	10	Ann Street Stop 10 at King George Square	-27.468003	153.02397	1	http://translink.com.a	0		
4	100	100	Parliament Stop 94A Margaret St	-27.473751	153.026745	1	http://translink.com.a	0		
5	1000	1000	Handford Rd at Songbird Way	-27.339069	153.043907	2	http://translink.com.a	0		
6	10000	10000	Balcara Ave near Allira Cr	-27.344106	153.024982	2	http://translink.com.a	0		
7	10001	10001	Nudgee Rd at Golf Course, stop 35/32	-27.372728	153.098237	2	http://translink.com.a	0		
8	10002	10002	Nudgee Rd at Golf Course, stop 32/35	-27.372787	153.098416	2	http://translink.com.a	0		
9	10003	10003	Approach Rd near Mellifont St, stop 31	-27.382612	153.090391	2	http://translink.com.a	0		
10	10005	10005	Handford Rd at Songbird Way	-27.339592	153.042985	2	http://translink.com.a	0		
11	10006	10006	Redwood St at Farrant Street	-27.402593	153.007388	2	http://translink.com.a	0		
12	10007	10007	Ridge St at St Vincents, stop 19	-27.385924	153.062459	2	http://translink.com.a	0		
13	10008	10008	Ridge St at Scott Street, stop 18	-27.387328	153.063647	2	http://translink.com.a	0		
14	10009	10009	Ridge St at Scott Street, stop 18	-27.386816	153.063368	2	http://translink.com.a	0		
15	1001	1001	Queen Street, 1E	-27.470552	153.024575	1	http://translink.com.a	0	place_QSBS	1E
16	10010	10010	Saint Vincents Rd near Ridge St, stop 19	-27.385663	153.061812	2	http://translink.com.a	0		

Figure 1: Sample csv Data

For this project, the main files were those containing stops, stop times, Go-Card© retailers, and data detailing the number of trips taken using public transport in a given year. All files were of a similar format to that seen in Figure 1.

Data Filtering and Preparation

Excel

The majority of low level filtering and data selection was done through Excel™, column/row highlighting capabilities were utilised to isolate data such as coordinates from the stops.txt file. Once the desired data was chosen, it was copied to another excel spreadsheet and saved directly as its own csv file. Beyond simple spreadsheet formatting, the most important utility for this project was the versatile Excel™ pivot table. This was used to find the total number of instances of a value within a data set, such as a bus stop ID or a suburb containing a Go-Card© retailer. Most of the data was processed and organised in Excel™ before being passed to programs like MATLAB™ for the visualisations.

Java

A small java program was written for the purposes of coordinate filtering. This involved reading the csv file in and then outputting the filtered results into a new file. The Java IDE Eclipse™ was used to both edit and run these written scripts.

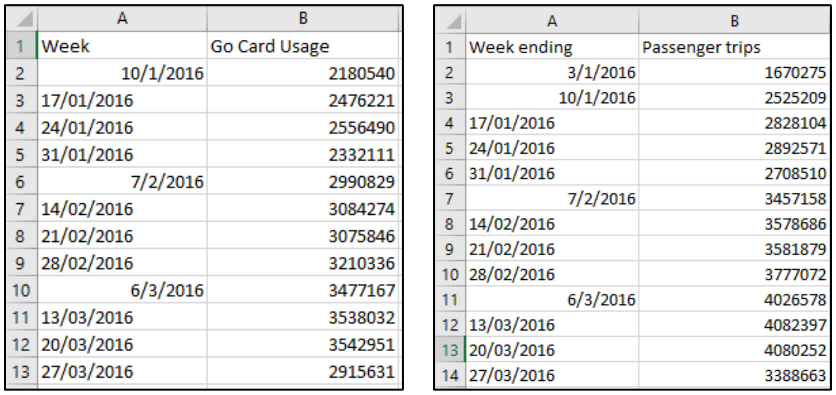
Visualisation Tools

Two software programs were used for this project's data visualisations: MATLAB™ and the QGIS™ Geographic Information System. Basic graphs and plots were run through MATLAB™ however more elaborate visuals such as those over a geographic region utilised QGIS™. The specific methods used for each visualisation are outlined with each generated figure.

Results

Yearly Transport Usage

A simple goal was to find out when are Translink™ services used the most, i.e. which month is the busiest time of year. Two data sets were combined for this, one containing the total number of Go-Card© trips and the other containing the number of trips total. Sample data is shown below:



	A	B
1	Week	Go Card Usage
2	10/1/2016	2180540
3	17/01/2016	2476221
4	24/01/2016	2556490
5	31/01/2016	2332111
6	7/2/2016	2990829
7	14/02/2016	3084274
8	21/02/2016	3075846
9	28/02/2016	3210336
10	6/3/2016	3477167
11	13/03/2016	3538032
12	20/03/2016	3542951
13	27/03/2016	2915631

	A	B
1	Week ending	Passenger trips
2	3/1/2016	1670275
3	10/1/2016	2525209
4	17/01/2016	2828104
5	24/01/2016	2892571
6	31/01/2016	2708510
7	7/2/2016	3457158
8	14/02/2016	3578686
9	21/02/2016	3581879
10	28/02/2016	3777072
11	6/3/2016	4026578
12	13/03/2016	4082397
13	20/03/2016	4080252
14	27/03/2016	3388663

The data was organised into weekly portions, and had entries going back to 2012. For this visualisation, the data retrieved in 2016 was used to ensure the result was the most up-to-date.

Figure 2: Go-Card Trips & Total Trips Data

Using MATLAB™, the two data sets were plotting on top of one another in the form of a bar graph, shown in Figure 3. The x axis specifies the general month and the y axis is the number of trips taken each week.

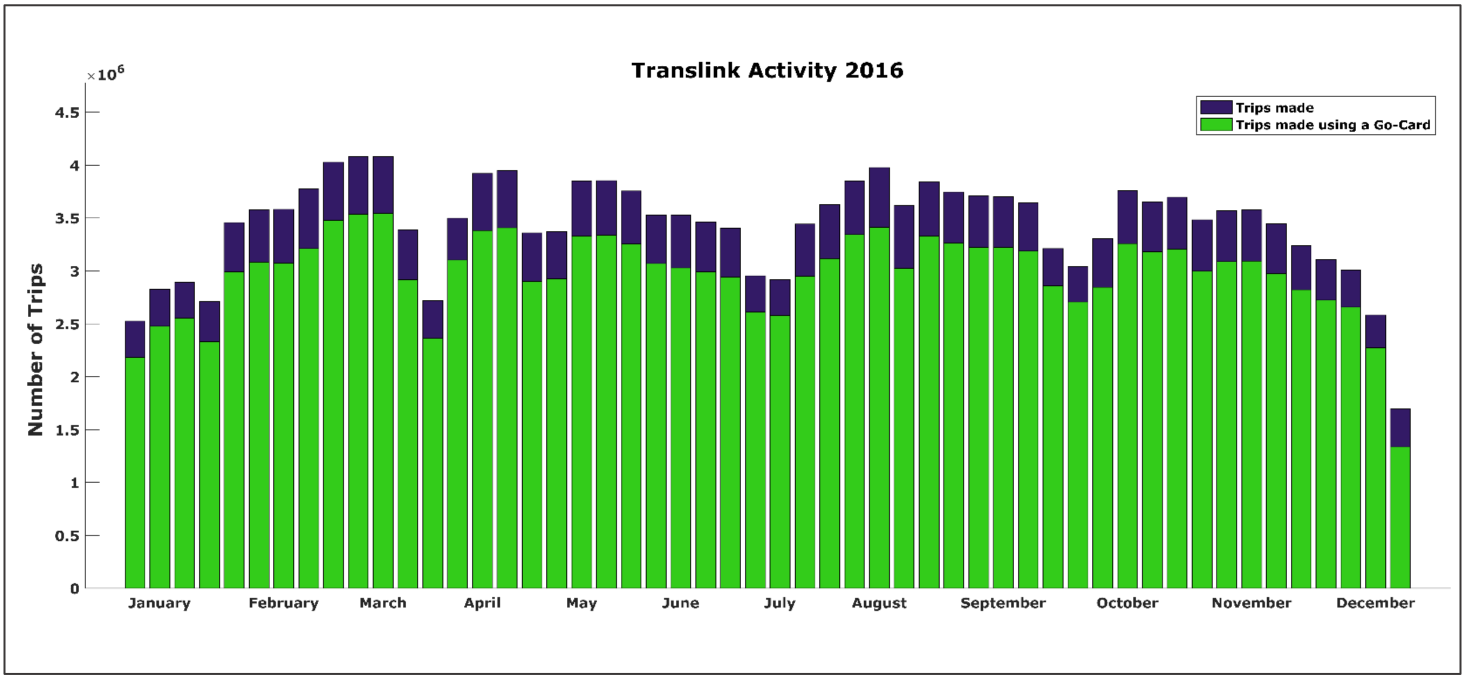


Figure 3: Translink Activity in 2016

A number of interesting trends can be seen here, most notably how the final week of December is the least active time of the year. This makes sense considering the Christmas holidays. Late February/early March is the most active time of year, perhaps because at this time of year <u>all forms</u> of schooling are active, including University. Three big dips are at even intervals, a likely explanation is they correlate to school holidays, signalling the end of each term. Regarding the graph itself, Go-Card© travel represents the vast majority of trips made, this is visualised explicitly in the simple pie graph shown in Figure 4.

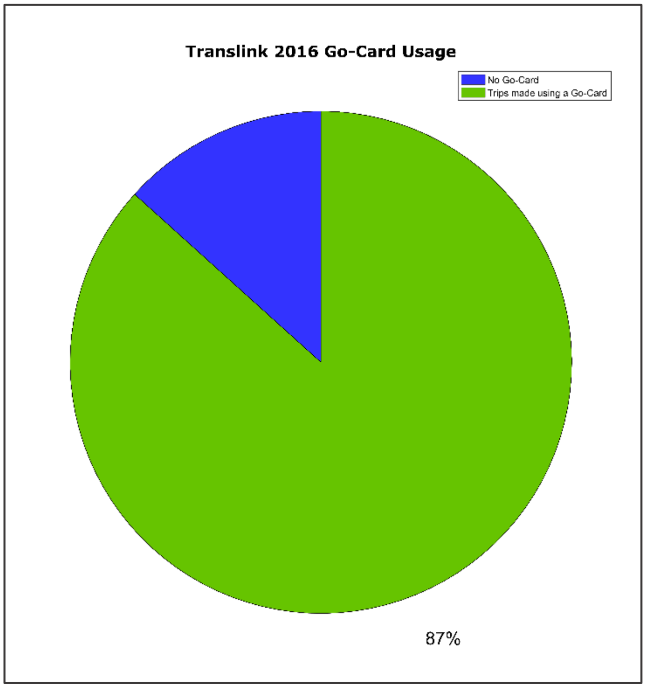


Figure 4: Percentage of Trips that use a Go-Card

Both 2016 data sets were summed and then the appropriate percentage values needed for Figure 4 were parsed into MATLAB™. The pie function was then used to generate the above figure. This chart is an indicator of both customer preference and crowd efficiency, 87% of all trips are made using Go-Cards©.

Go-Card© Retailers

One csv contained the information of every registered retailer that sells Go-Cards© to the general public. It contained both coordinate data and address information. The address information was run through a pivot table to count the occurrences of each suburb. This was then plotted in MATLAB to create Figure 5. This reveals how Brisbane Central is the most concentrated area for Translink™ retailers.

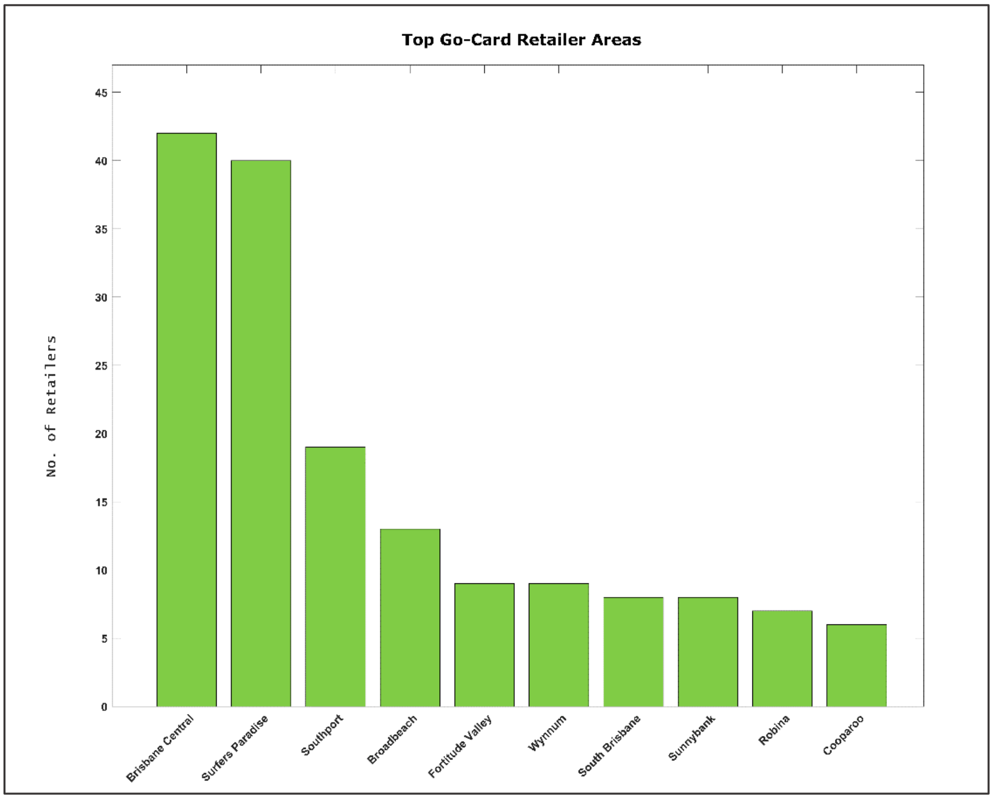
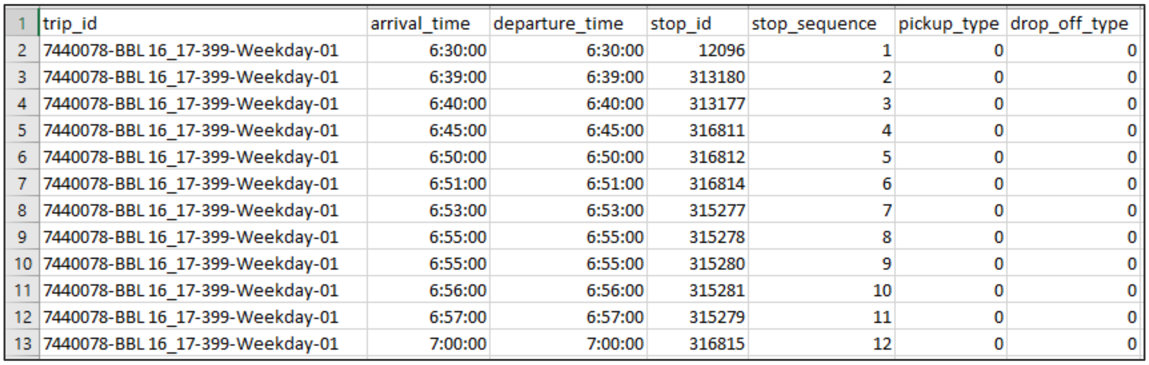


Figure 5: Top 10 Go-Card Retailers

Daily Translink™ Activity

The next visualisation was made to deduce what time of the day public transport is most active, not over a week or year as was done before. This used the largest file downloaded, the file containing stop times. Every row represents a single instance of a bus, train or ferry arriving at a station. The file contained over one million lines, a sample is shown below:



1	trip_id	arrival_time	departure_time	stop_id	stop_sequence	pickup_type	drop_off_type
2	7440078-BBL 16_17-399-Weekday-01	6:30:00	6:30:00	12096	1	0	0
3	7440078-BBL 16_17-399-Weekday-01	6:39:00	6:39:00	313180	2	0	0
4	7440078-BBL 16_17-399-Weekday-01	6:40:00	6:40:00	313177	3	0	0
5	7440078-BBL 16_17-399-Weekday-01	6:45:00	6:45:00	316811	4	0	0
6	7440078-BBL 16_17-399-Weekday-01	6:50:00	6:50:00	316812	5	0	0
7	7440078-BBL 16_17-399-Weekday-01	6:51:00	6:51:00	316814	6	0	0
8	7440078-BBL 16_17-399-Weekday-01	6:53:00	6:53:00	315277	7	0	0
9	7440078-BBL 16_17-399-Weekday-01	6:55:00	6:55:00	315278	8	0	0
10	7440078-BBL 16_17-399-Weekday-01	6:55:00	6:55:00	315280	9	0	0
11	7440078-BBL 16_17-399-Weekday-01	6:56:00	6:56:00	315281	10	0	0
12	7440078-BBL 16_17-399-Weekday-01	6:57:00	6:57:00	315279	11	0	0
13	7440078-BBL 16_17-399-Weekday-01	7:00:00	7:00:00	316815	12	0	0

Figure 6: stop_times.txt sample data

The arrival time specifically was the key to this visual. This column was isolated and taken to another spread sheet. Using the Excel™ Pivot Table function, every arrival time was counted and sorted over 24 hours. These tallies for each hour can serve as an "activity index", e.g. between 0700 and 0759, x amount of services stopped at locations across SE Queensland. Once tabulated, the data was passed as a csv to MATLAB™ to produce Figure 7 below, which uses an area plot instead of a bar plot.

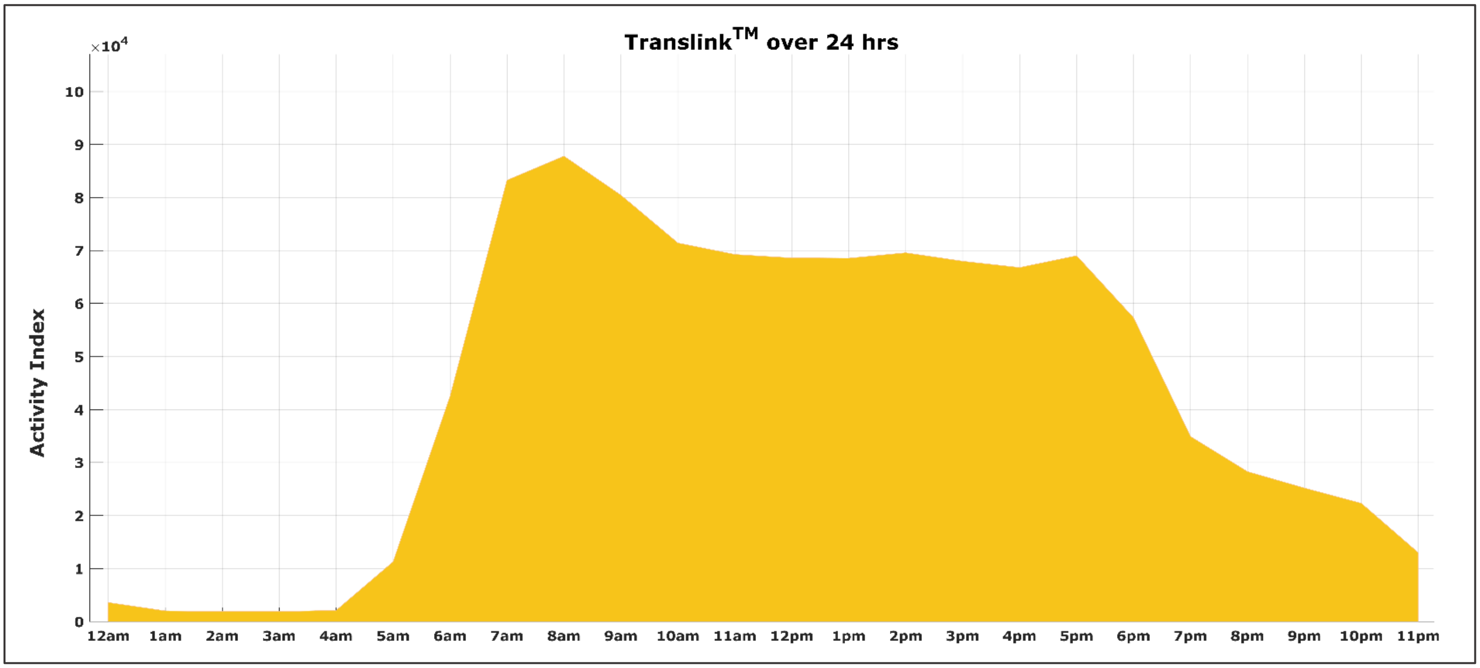


Figure 7: Average Daily Translink Activity

As shown above, the hours when people are commuting IN for work is much busier then when people are heading home. A huge amount of activity begins at 6am, continuing until 10am. One would have expected a dip in activity during the midday/afternoon hours, but instead, Translink™ services visibly plateau until ~6pm. Between 1am and 4 am, very, very, few services are active, an expected result seeing as the majority of the population is asleep at this time.

Distribution/Density of Stops

The file containing stop information showed the promise of generating a map that could reveal stop distribution. Containing over 12000 lines, the data set has coordinates for every single stop across SE Queensland. Figure 1 in the Methods section gave a sample of this. The latitude and longitude values were lifted from the spreadsheet and placed into their own file. Once isolated, they were parsed into MATLAB™.

A number of possible functions were looked into, but in the end the hist3 function did exactly what was needed. The hist3 function is used for constructing bivariate histograms. By treating the latitude and longitude as independent univariate data, the function visualizes crosstabulations in the two variables. This of course is guaranteed to happen because they are discrete coordinates. Figure 8 shows this from a 3D perspective. For this figure, the axes had to be appropriately scaled to ensure geographic continuity and the number of data bins for the histogram had be high to ensure sufficient detail. A custom colour scheme was also made to better distinguish non-zero values.

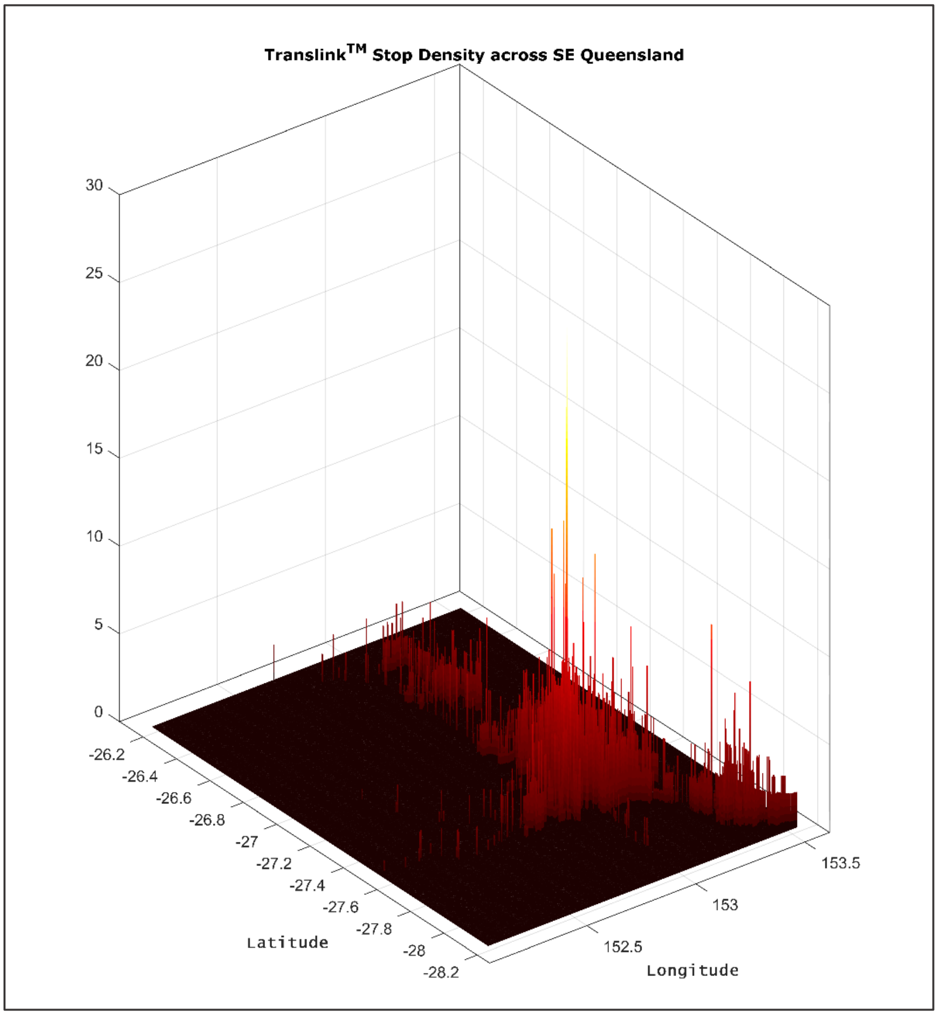


Figure 8: Coordinate Data using hist3

This demonstrates what a bivariate histogram produces, and, in this context, how the calculated z value directly represents the density of stops in a given area. While this plot is noteworthy, it is still rather hard to infer information from it. To fix this, the view perspective was changed to a top-down view (view(2) in MATLAB™) and a colour bar was added to the side of the figure. This, ultimately, resulted a 2D heat map of stop densities (Figure 9, next page).

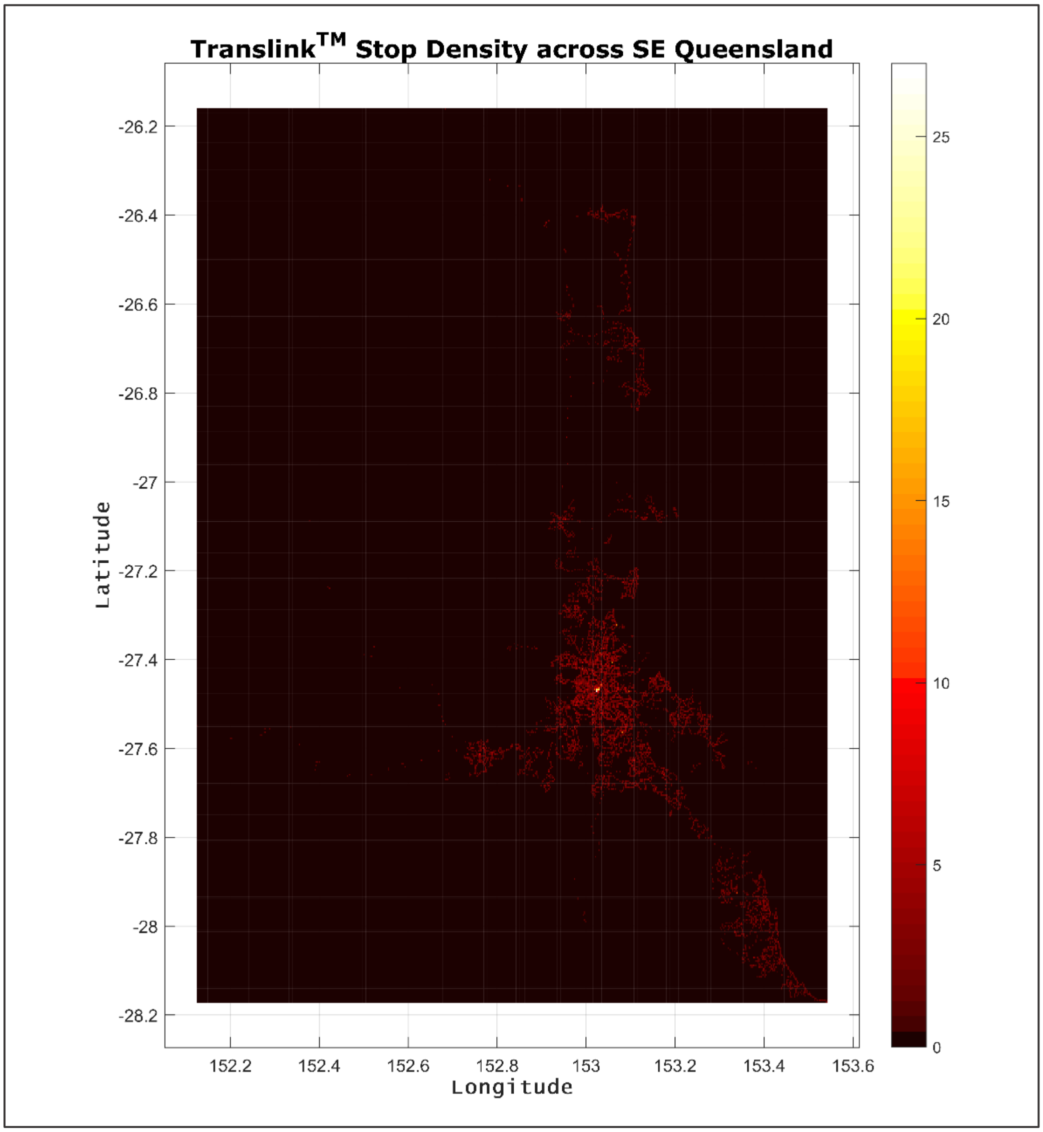


Figure 9: Bivariate Histogram Stop Map

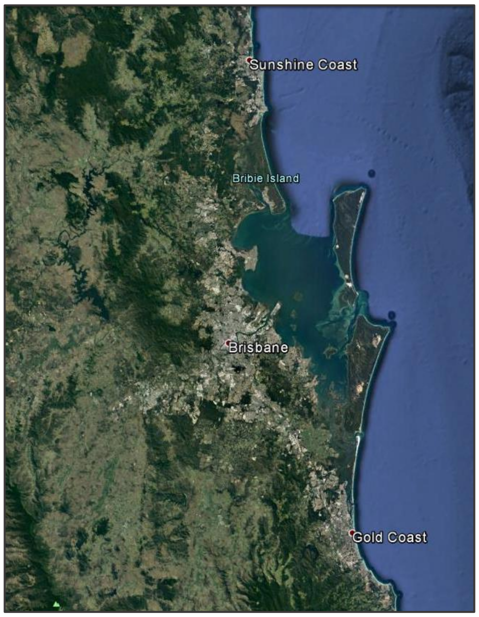


Figure 10: Google Earth Perspective

From this view in Figure 9, much more detail and information can be derived from the plot. To those familiar with Queensland geography, notable regions and landmarks might already be easy to spot. It is fascinating that by using only stop coordinate data, all major urban areas of southeast Queensland have been mapped. The bright spot represents the centre of Brisbane, along with its surrounding suburbs. Ipswich, the sunshine coast and the gold coast can be seen too. The mini-map made using Google Earth (Figure 10) supports this result.

Figure 9 provides a general overview of SE Queensland. To get a more detailed view of the Brisbane area, the coordinate data must be sorted and run through the same script (to ensure a detailed plot). A simple Java program was written to sort the data into a new file, which would 'zoom' in on a selected coordinate region whilst maintaining the relative order of the data. This was done twice, once to plot Brisbane and again to plot the inner suburbs. Going down to this level of detail reveals the true discrete nature of the data, stops dot their way through major roads and around geographic areas (next page).

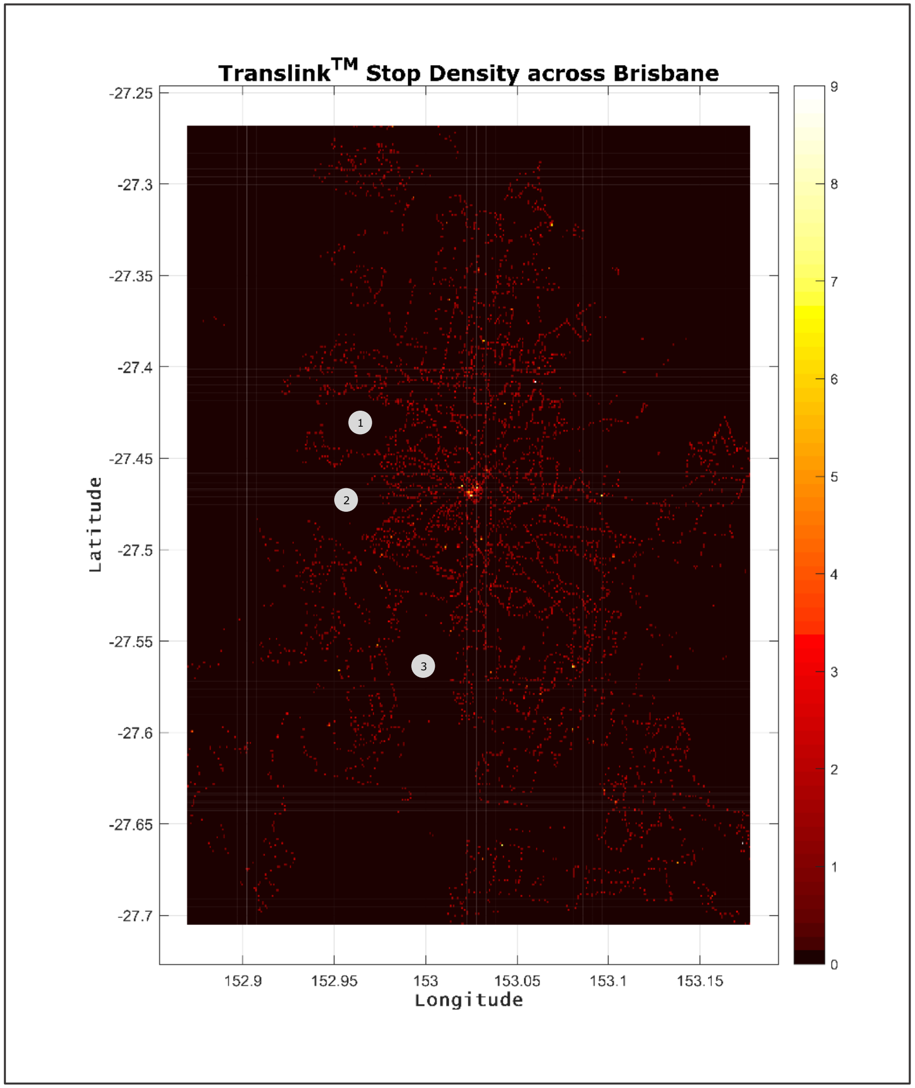


Figure 11: Stop Density in Brisbane

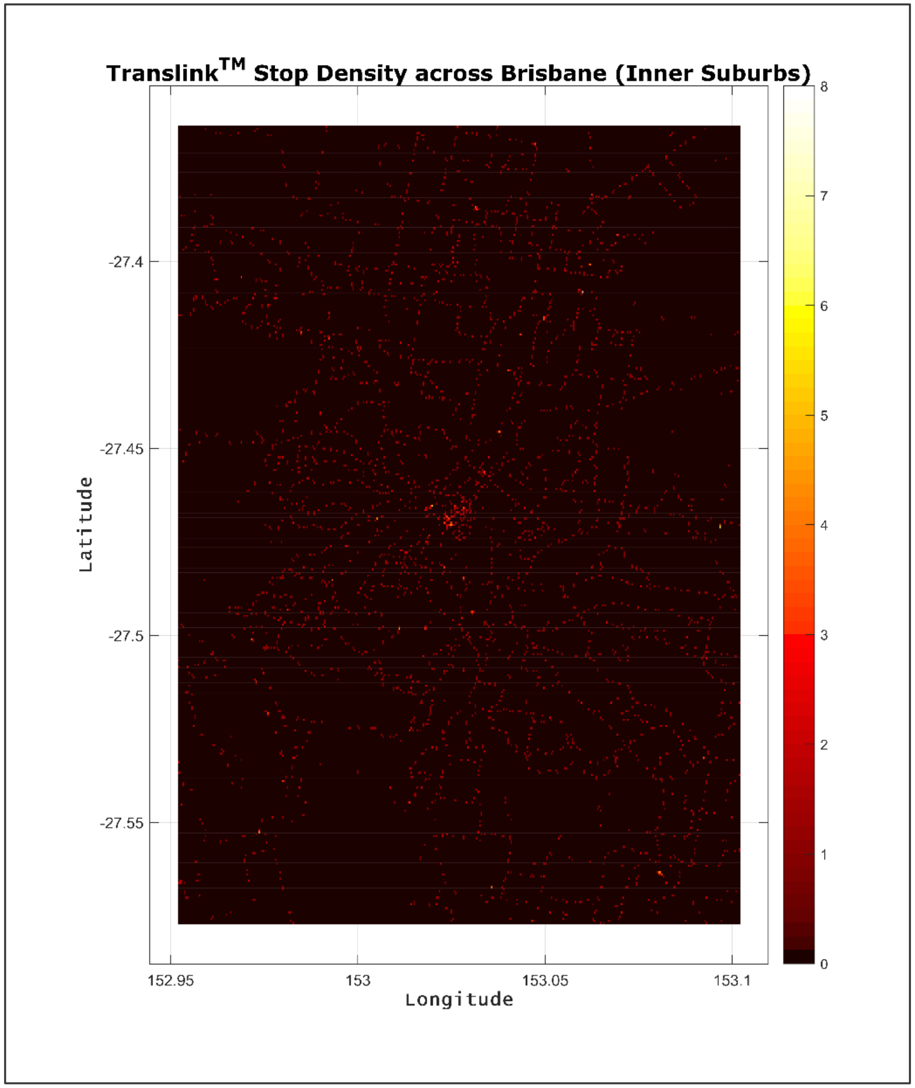


Figure 12: Stop Density (Inner Suburbs)

At each stage of "zooming" the level of detail is noticeably increased, down to the street level. Translink™ stops snake their way through and around Brisbane suburbs. Unsurprisingly, voids in the plot represent geographic landmarks. For example, the numbered markers 1, 2 and 3 on Figure 11 show the Enoggera Barracks, Mt Cootha and Archerfield Airport respectively. Both of these plots show that Translink™ stops are not in short supply, especially in the Brisbane region. The bigger question remains, how active are each of these stops. The next section explores this topic.

Stop Efficiency/Effectiveness

To depict data as a heatmap, a larger and more graphically capable piece of software was downloaded and used: QGIS. This is a Geographic mapping tool offering hundreds of plugins for mapping data and creating visualisations. Figure 13 is an example of what QGIS can create. The same data from the previous section was imported into the current project and a map overlay was aligned using the *OpenStreetMaps* plugin. The result is more continuous and flowing plot then like the granular/discrete plots seen previously. By manipulating the properties of the coordinate data layer, the heatmap setting was selected and tweaked to produce Figure 13.

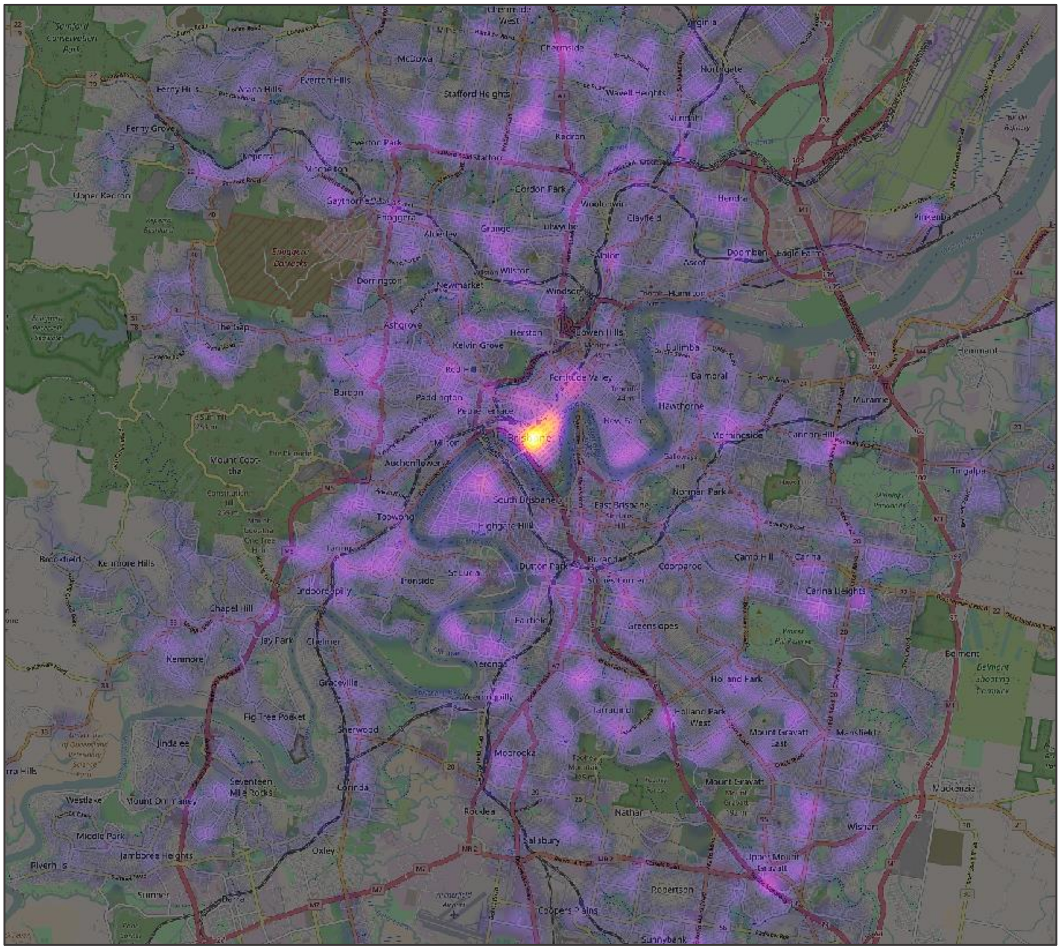


Figure 11: Stop density as a heatmap

The heatmap again shows how the CBD contains the highest stop density across SE Queensland. This figure demonstrates QGIS's heatmap capabilities. The next crucial step is to find a measure of stop activity, i.e. find a z 'intensity' value for the already established x y coordinate values.

Luckily, this process involved a near identical approach to how Figure 7 was created (Translink™ activity over 24 hours). The stop times file containing over a million lines gave both times for each occurrence AND the stop ID of the stop where the instance took place. Instead of using the Excel™ pivot table to count the occurrences of each time slot, it can be switched to count stop ID's instead. By matching these stop ID's with their relative coordinate position (located in the stops file), the appropriate 'activity index' can now be assigned to each x and y coordinate pair. This third value of 'activity' is what is needed to create the heatmap of the most active public transport locations/areas in Brisbane, Figure 14 (next page).

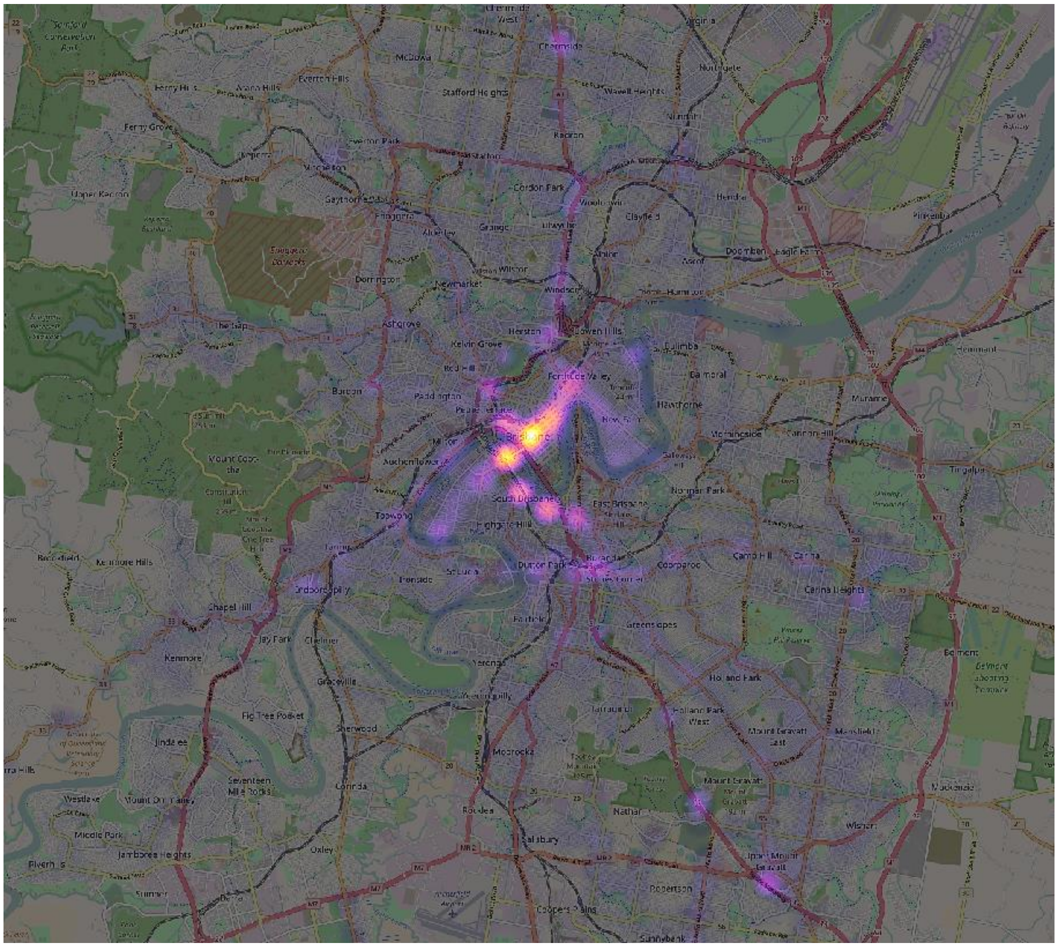


Figure 12: Active Transport Areas of Brisbane

Figure 14 paints a very different picture in comparison to Figure 13. Stop activity suddenly becomes very localised to the inner city (seen in greater detail using Figure 15) and in small pockets like Chermside and Mt Gravatt. Each of these hot spots represent places that, over 24 hours, process the most buses and, subsequently, the most bus routes. This is a very important result because it shows that even though areas like Mt Gravatt are far away from the CBD, there still exists a public transport 'hub' which can be used for efficient daily commuting.

Ultimately, the majority of bus routes and train lines have to converge somewhere, and with that will come significantly higher levels of activity. Figure 15 is fascinating in how it pinpoints these high-use stops. It is easy to recognise locations like UQ lakes, Southbank Bus Station, Roma Street and of course, Cultural Centre Stations 1 and 2. Platform 1 at the Cultural Centre ended up having the highest activity index, followed closely by Platform 2.

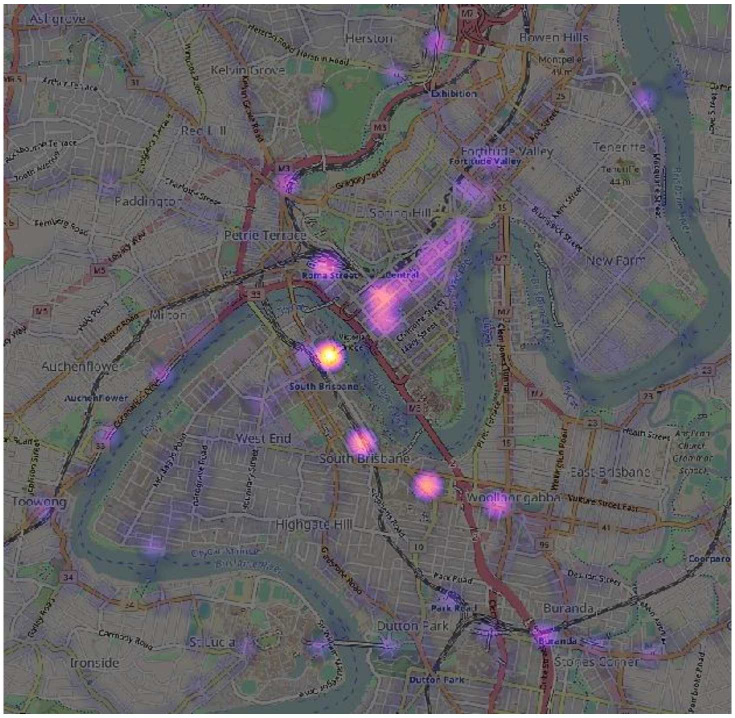


Figure 13: Active Transport Areas (Inner City)

Conclusion

Trends

Overall, this project succeeded in creating the proposed data visualisations. Each of the generated results provided useful information and the trends identified were reaffirmed by considering both the real world and personal experiences with the Translink™ agency.

The first section saw periodic patterns in how often public transport is used across SE Queensland. The time of year correlated to either a rise or fall in Translink™ usage, whether it was the Summer holidays or the end of a school term when students ceased commuting during the week. Investigating this further revealed, on average, when public transport in SE Queensland is most active over a single day. The morning hours between 1am and 4am saw a distinct drop in activity. Providing more services during these hours could possibly be beneficial to individuals currently working under adjusted night-time schedules.

Plotting stop density through MATLAB™ produced some visually stunning results. Urban development and high population can be directly correlated to public transport services being implemented and subsequently, stops being constructed. The efficiency/overall activity of individual stops revealed less obvious patterns, most notably how some suburbs such as Indooroopilly, Chermside and Mt Gravatt are highly active areas despite being far from the CBD. This showed that some regions are more active, and thus provide possibly smoother commutes than neighbouring suburbs.

Future Improvements /Visualisations

Combining the produced visualisations with population and census data across Brisbane was one of this project's initial intent. However, over the course of the project it was difficult to obtain data of a similar format and more importantly, data that was as granular and detailed as the Translink™ data. Extending these visualisations to include localised census data could reveal some very useful correlations. Specifically, if suburbs with a high population are somehow lacking in transport services, indicating the need for new Translink™ bus or rail routes to be made.

Similar visuals could be made using data from other cities across Australia, revealing which cities have the most extensive transport system/are the most efficient. Trends in efficiency could then be identified and applied to better improve public transport across the country.

References

Code Used to create Figures (MATLAB™ and Java)

https://github.com/ryanjphelan/COSC3000-2017-Data-Visualisation

Sources

*Daniel Hurst (*20 May 2011*). "Bus Overcrowding Worse than Ever".* Brisbane Times. Fairfax Media*.* Retrieved 20th April 2017*.*

Open Data. (n.d.). Retrieved April 3rd, 2017, from

Software

MATLAB Release 2016b, The Math Works, Inc., Natick, Massachusetts, United States.

QGIS Geographic Information System (2009). QGIS Development Team. Open Source Geospatial Foundation. http://qgis.osgeo.org