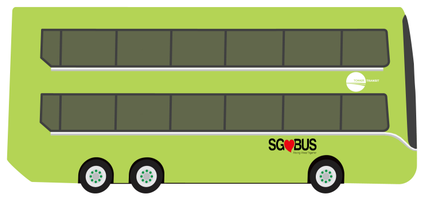
GUIDE TO SINGAPORE'S BUS COMMUTER VOLUME

VISUAL ANALYTICS PROJECT

ISS608
AY2022-23

TEAM



AUTHORS

AISHWARYA SANJAY MALOO
HUO DA
PRACHI RAJENDRA ASHANI

MENTOR

PROFESSOR KAM TIN SEONG

BACKGROUND & OBJECTIVE

Singapore public transport system serve 5 million daily commuters. Historically ranked amongst the world's best, concerns around travel times and overall reliability still exist. These issues have been exacerbated by an increasingly growing population, necessitating the need for it to consistently adapt to consumer patterns.

The primary goal of this study is to identify patterns in the commuter flow across subzones across different hours of the day, over weekday and weekends/holidays, and over three-month period – October, November, and December 2023. For this we perform exploratory data analysis to draw insights about commuter distribution and density. Our second objective is to perform a comparative analysis of various Spatial Interaction Models and assess their viability to predict commuter flow in the future. Such models also help identify explanatory variables that impact Singapore's bus ridership at the subzone level. Our third objective is to perform clustering analysis using different methods. Clustering analysis can help tailor public bus services per the requirements of identified cluster of commuters.

Such a data driven analysis can help the Land Transport Authority and other government public service organizations to identify problem areas in the public bus services, make enhancements to better suit the requirements of Singapore populace, as well as make the services more time and cost efficient.

DATASET AND METHODOLOGY

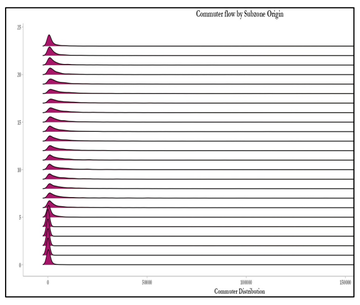
The datasets used in this project were provided at LTA Data Mall. For this project, we obtained the passenger commuting data in 3 month, i.e. Oct 2022 to Dec 2022. Each trip's information describes its original bus stop, destination bus stop, passenger's entry timestamp (in hour) and whether on weekdays or weekends/holidays. Moreover, we have 5 other datasets containing information on each bus stop, each subzone, each planning area, each region of Singapore as well as the distance between each subzone. Four of the data sets are spatial data sets. The data was wrangled using R and several new dataframes were created which have been put up in our application.

We used R Shiny to create relevant interactive charts, compelling visualisations and modelling & analysis such as clustering and regression so that the users can have a better understanding of Singapore's commuter volume.

The libraries used for data wrangling were rgdal, spdep, tmap, sf, ggpubr, cluster, factoextra, NbClust, heatmaply, corrplot, psych, Hmisc, knitr, kableExtra, ClustGeo, ggiraphExtra, plotly, ggstatsplot, tools, scales, extrafont, ggridges, gganimate, viridis, caret, gtsummary, gt, treemap, shiny, shinycssloaders, shinydashboard, shinythemes, tidyverse

EXPLORATORY DATA ANALYSIS

Analyzing commuter distribution



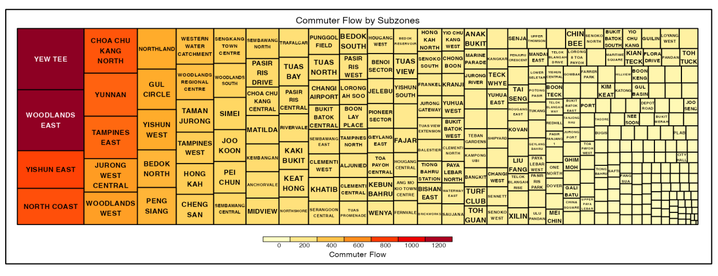
Distribution at origin by time

Using ridgeline plot, we analyze the distribution of commuters in from both origin and destination perspective for features such as month (Oct, Nov, Dec), day type (weekday, weekend), time of day (12 AM to 11.59 PM), subzone, planning area, and region.

We observe a significant right skew in the commuter volume from both origin and destination perspective. This indicates a very high asymmetric flow of commuters using the bus services.

Analyzing commuter density

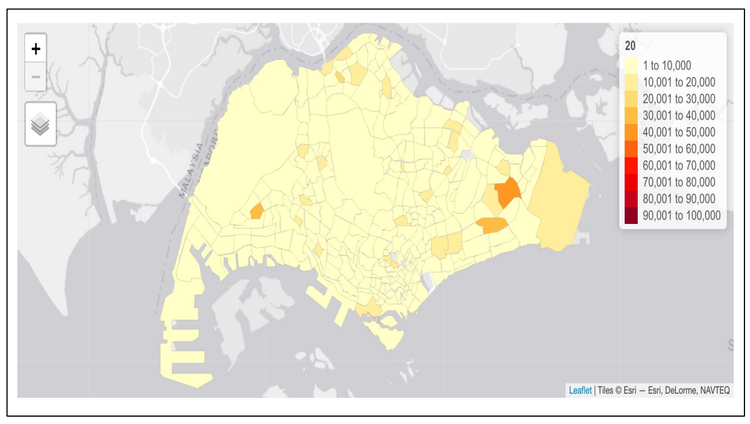
We see high commuter density in originating from Yew Tee, Woodlands East, Yishun East, and North Coast subzones. One reason that can be attributed to such high density is that these regions are situated on the east and the west of Singapore and primarily dominated by residential areas. Residents in these areas would be using bus services to travel to their workplaces, and education institutes on the weekdays and recreational areas over the weekends.



Density of commuters by at origin subzone

SPATIAL COMMUTER FLOW MAP

Analyzing number of commuters in subzone



A choropleth map has the advantage of providing an easy-to-understand display of data that would be difficult to explain using other kinds of visualization, such as tables or graphs. The map illustrates the relative disparities between areas by giving colours or tones to distinct values.

we use choropleth map to assess the number of commuters in a subzone. Users can choose day type (weekday, weekend/holidays) and time of day (12 am to 11.59 pm) to assess the number of commuters in a subzone.

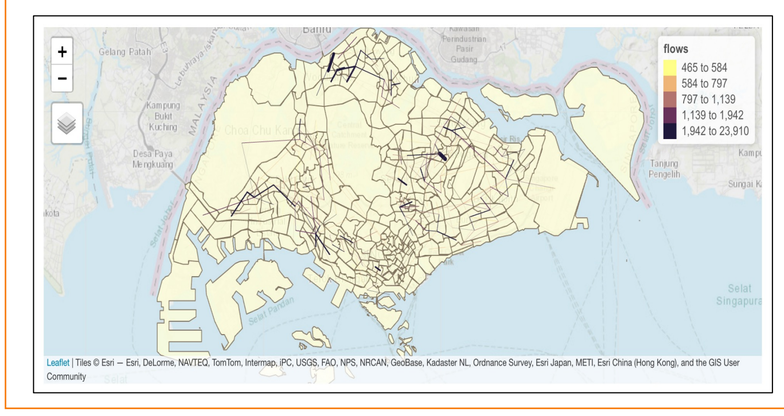
Choropleth when appended with traffic tracking features can be extremely useful to assess the density of commuters in a subzone in real time. Such real time assessment can be necessary to take necessary actions to smoothly ease the passenger traffic in buses.

FLOW MAP

An origin-destination(OD) flow map depicts the movement of people, goods, or other entities from their origin to destination. An OD flow map's primary aim is to provide a visual representation of the volume and direction of movement between sites.

In transportation planning and logistics, OD flow maps are often used to identify patterns and trends in travel or shipment behavior. An OD flow map, for example, might assist a city or regional planner in understanding commuting trends between subzones/regions/cities. It can also be used to locate bottlenecks or locations of congestion in transportation networks.

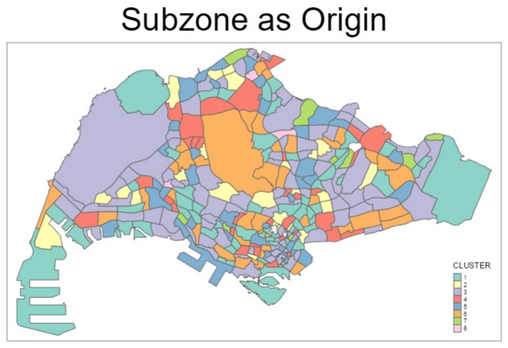
In the context of Singapore, this can very useful given the current climate. The public's perception of the reliability of public transport has been falling down, albeit by 5%. This map can be used by urban planners in Singapore to assess the commuter volume by subzones. On the basis of this, they can add bus stops where the demand/commuter volume is high, remove redundant bus stops where the commuter volume is low. They can also do a time analysis with the commuter flow to see peak times and subzones.



CLUSTERING ANALYSIS

Understanding commuter patterns and behaviours through clustering

Subzone as Origin

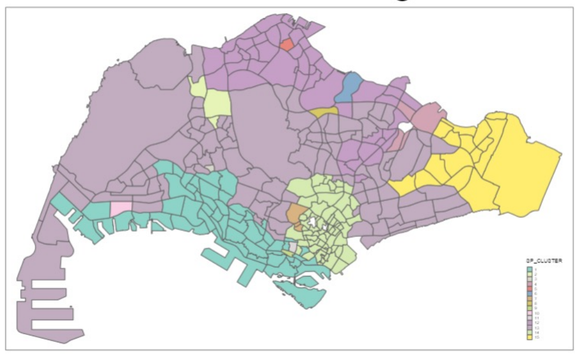


Hierarchical cluster for Origin Subzones

We use day type (weekdays, weekends/holidays) to perform hierarchical clustering. The users can choose weekdays, weekends/holidays, or both to run this clustering method. The user can choose the number of clusters per their requirement to derive passenger clusters.

A detailed study on the factors influencing cluster formations can help understand help the authorities to better understand the commuters and their requirements better and correspondingly optimize flow in areas that are have high commuter volume.

Subzone as Origin



Geospatial constrained SKATER cluster for Origin Subzones

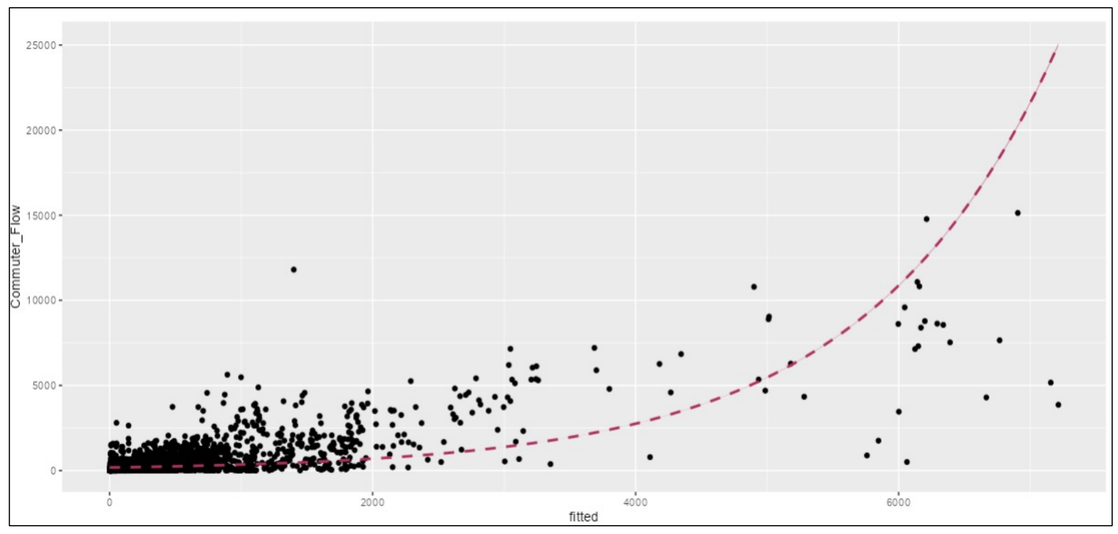
Like hierarchical clustering, in geospatial constrained SKATER cluster analysis, the user can choose weekdays, weekends/holidays, or both to run this clustering method. The user can choose the number of clusters per their requirement to derive passenger clusters. Furthermore, the users can also choose clustering type – Euclidean, Maximum, Manhattan, Canberra, Binary, Minkowski, and Mahalanobis. Different clustering techniques can help in making better statistical inferences.

REGRESSION ANALYSIS

As flow is non-negative integer counts, bus ridership was fitted using generalized linear model with Poisson Distribution, with a logarithmic link to transform the independent variables, rather than a tradition log-normal regression model.

We perform unconstrained, origin-constrained, destination-constrained, and doubly-constrained generalized linear mixture regression.

We observe that the doubly-constrained model has the best R^2 score, RMSE, and MAE compared to the other models. All input variables are statistically significant (p-values < 0.05). However, with around 60% R^2 score, there is room for enhancement for doubly-constrained spatial interaction model. The input variables include – origin subzone, destination subzone, subzone distance, ……



Regression plot for commuter flow for doubly constrained model

Model	RMSE	R^2 Score	MAE
Unconstrained	573	0.54	237
Origin Constrained	573	0.54	237
Destination Constrained	511	0.63	212
Double Constrained	511	0.63	212

NEXT STEPS...

1. Currently, the doubly constrained generalized linear mixture model has an R^2 score of 60% and a high AIC score. we plan to enhance the by increasing input features such as median income at destination and origin subzone, proportion of type of dwellings at the subzone, number of bus stops at the subzones, to name a few. This can likely help to improve the prediction accuracy of the model.
2. We also plan to develop other models such as gaussian distribution models and gravity model. this will help to draw insights from model comparison, and find the best model that can help to achieve the project objectives.
3. We plan to deep dive into individual clusters resulting from clustering analysis. Studying these clusters in detail is essential for identifying commuter needs and behaviour.
4. We plan to harness better commuting power. Currently, a proportion of data is being used through random sampling to perform supervised and unsupervised machine learning modelling. Better commuting power can allow inputting large samples which can in turn improve model analysis.