Meta Data of the SA2 level origin-destination flow estimation

August 28, 2023

Contents

Chapter 1

Motivation

Understanding and analysing human mobility data is crucial for various sectors, including transportation, traffic management, ride-sharing, logistics, urban planning, social computing, disaster and emergency response, health informatics, and pandemic prevention [Mokbel et al., 2023, Xu et al., 2023, Song et al., 2016]. Utilizing mobility data to understand human behaviour can lead to effective solutions to traffic management problems. Understanding how, where, and why people move to cities helps understand the infrastructure and energy demand, reduces urban inequalities, and improves urban safety, and situational awareness, enabling robust infrastructure and protecting cities from disasters like forest fires, earthquakes, and emergency management [Barbosa et al., 2018]. Statistical areas are a prime focus for mobility applications, introducing various mobility modalities like public transport and private vehicles e.g., electric vehicles, bicycles, and scooters with sharing programs considering traffic congestion is a global issue, with drivers spending 6.9 billion driving hours and wasting over 11 billion litres of fuel per year [Mokbel et al., 2023]. Monitoring and reducing emissions are challenging when data are collected from in-situ sensors. In such contexts, mobility data can help estimate relevant data to understand the effects of e-mobility, collective transportation, and infrastructure improvements in emission control [Mokbel et al., 2023]. Clearly, mobility between statistical areas is a critical aspect for understanding human movement patterns and essential for efficient urban planning and transportation management [Gonzalez et al., 2008].

Extensive geolocated datasets have enabled the quantitative study of human movement patterns, enabling scientists to generate models that capture and reproduce spatiotemporal structures and regularities in human trajectories.[Barbosa et al., 2018]. Personal digital devices, like mobile phones, significantly increase human mobility data availability, enhancing research in activity recognition, personalized routing, and crowdsourcing in human mobility analysis [Matekenya et al., 2021]. Human mobility research presents opportunities and challenges in developing accurate activity models in context-aware settings, considering multi-modal transportation networks. Challenges include data availability, quality, privacy, bias, low incentives for data sharing, and the right simulation approach for synthetic data generation. Large-scale aggregated datasets can help in high-level analysis but are too small to understand, analyze, and predict human behaviour [Mokbel et al., 2023].

This document contains the technical and operational metadata that is required to estimate origin-

destination (OD) movement flow among a set of geographic regions, connected by either public transport or private vehicular network. This metadata describes a thorough methodology for producing mobility data assets for various statistical levels in Australia, with a focus on OD flow matrices at the Statistical Area 2 (SA2) level. The number of movements between pairs of SA2 regions that take place over predetermined time periods, such as weekly or monthly intervals for a full year, is represented by these matrices. Both private automobiles on the road and public transportation (such as buses, trains, trams, or ferries) are taken into account while gathering data.

The following are the key aspects of discussion while creating the output for each state:

- Input data collection: Utilize publicly available heterogeneous and multi-modal human mobility datasets to gather information about people's movements across different SA2 regions in Australia. Detailed explanations of the data sources used, including various publicly available datasets that provide insights into human mobility across Australia.
- Methodology: A step-by-step methodology is provided to generate the SA2 level OD flow matrices, which may involve data preprocessing, aggregation, and analysis. The methodology used is described in detail to ensure the reproducibility of the work for future reference and research prospects, including information on software, tools, and libraries used in the analysis, as well as any necessary code or scripts to replicate the results.
- Output OD-flow matrices: Create OD flow matrices at the SA2 level, which indicate the number of movements between each pair of regions for specific time intervals.
- Illustrations: Visualizations of the generated mobility data are provided whenever needed by showcasing patterns and trends in human movement at different statistical levels.
- Time intervals: Consider both weekly and monthly time intervals for data collection to capture variations in mobility patterns.
- Modes of travel: Categorize movements based on the mode of travel, distinguishing between public transport and private vehicles on the road.

Overall, this document serves as a valuable resource for researchers and policymakers interested in understanding and analyzing human mobility patterns within Australia. It could be used for various purposes, such as urban planning, transportation management, and policy development related to public and private transportation.

Chapter 2

QLD SA2-level OD matrices

2.1 Introduction

This chapter contains the technical and operational metadata that is required to estimate OD movement flow among the SA2 regions in NSW, connected by PT networks. The Statistical Area 2 (SA2) regions of Queensland connected by buses, trains, trams and ferries have been used to evaluate OD movement flows. The passenger OD movement data among different stations(or the station-based OD flow) are first estimated using a statistical estimation methodology. The stations-based OD flow data are then translated into region-based OD matrices using the state-of-art method. Below a brief description of the method is provided along with the input-output data, the assumptions and the future work that remains.

2.2 Input data

Queensland has four years of public transport data, buses, trains, trams and ferries. Each SA2 contains multiple PT stops of each category and multiple SA1s. The input data typically contains the following columns for each month of the year:

1. *operator* - This column describes the operator name, for example, Surfside Buslines, Queensland Rail etc.
2. *month* - This column describes the data collection month of a specific year.
3. *route* - This column describes the unique travel route number in the bus PT network, and specifies the mode of transport for rail, tram or ferries.
4. *direction* - This column describes the inbound, outbound, north, south, east, west, clockwise etc. for buses, and rail for trains.
5. *time* - This column value is either weekday or weekend, depending on the data collection time.
6. *ticket_type* - This column provides the description of ticket type, either paper or go card.

7. *origin_stop* - This column provides the unique origin stop id of a journey.
8. *destination_stop* - This column provides the unique destination stop id of a journey.
9. *quantity* - This column provides the number of passengers travelling in the OD pair.

Figure 2.1 is illustrating geospatial distributions of PT stops of different categories along with the SA2 boundaries in black polygon and the SA2 centroids in grey star marker.

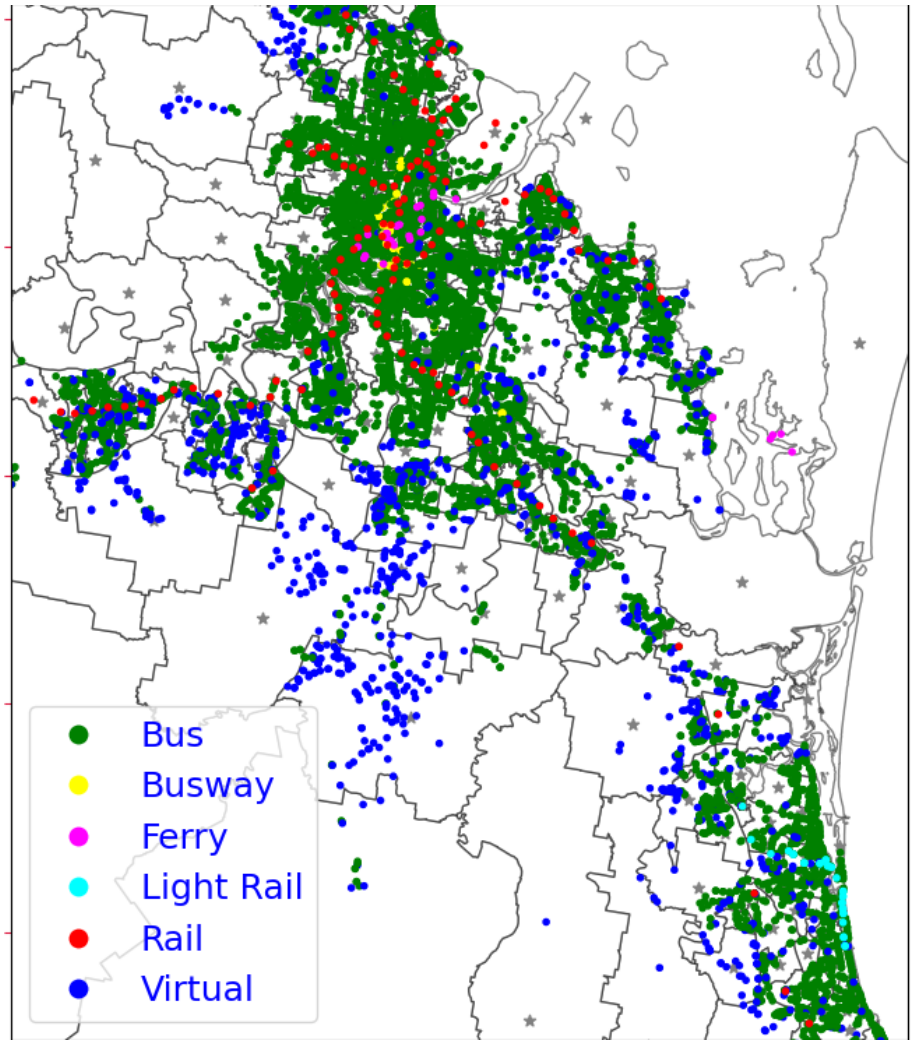

Figure 2.1: PT stops and SA2 boundary of Queensland along with centroids in grey is shown in different colours.

2.3 Method

The preprocessed input data contains station-based OD flow among each pair of stations of a particular mobility network. The next step is to assign probabilities of a population from an origin SA2 to boarding from an origin station to reach a destination station which is assigned to another probability of being part of a destination SA2.

2.3.1 Assumptions

This is achieved by the following assumptions:

1. Filtering out all possible PT stops for each SA1 centroid based on ABS (Australian Bureau of Statistics) census data. The assumption is a person is likely to travel within a radius given by a specific mean and standard deviation from the home location. It is assumed that people will not travel beyond the mean and 3-standard deviation from the SA1 centroid of interest to board a desirable public transport.
2. Assigning each PT stop to an SA1. If one PT stop is shared by multiple SA1, a population-based probability is assigned to SA1 to the PT stop.
3. The aggregated SA1-based population multiplied by the assigned probabilities should sum up to the station-based OD flow and hold true for the SA1-based movement for each SA1 within the SA2.

Next, the SA2-based OD matrix is derived, based on the probability of route choice and origin-destination PT stops. The following algorithm is used to select the nearest PT stop for each SA1 centroid.

- For each SA2, the t-value is computed using the mean and 3 standard deviation of distance to work census data.
- For each SA2, t value is used as a diameter to draw a circle centring the SA2 centroid to select all stations within this circle.
- For each SA1 of an SA2, the closest transit point is selected to be the SA1 centroid limiting the search to transit points selected in step 2.

After this step, each SA1 is mapped to one PT stop of each category.

2.3.2 SA2 based OD flow using population-based probabilities

Figure 2.2 is illustrating a set of the QLD PT stops with red, blue, green, light blue, lime, and pink markers with SA1 boundaries and centroids in grey. It is evident that each SA1 can contain one, multiple or no PT stops depending on the geographic distributions.

Let $SA2_O$ be the origin and $SA2_D$ be the destination. $SA2_O$ contains N_O number of SA1s with n_o number of distinct PT stops, and $SA2_D$ contains N_D number of SA1s with n_d number of distinct PT stops.

$\sum_{i=1}^{n_o} p^O(i) = 1$ where $p^O(i)$ is the probability of i^{th} PT stops being boarded from origin SA2 $SA2_O$. Because $N_O \geq n_o$, $\frac{1}{N_O} \leq p^O(i) \leq \frac{n_o}{N_O}$.

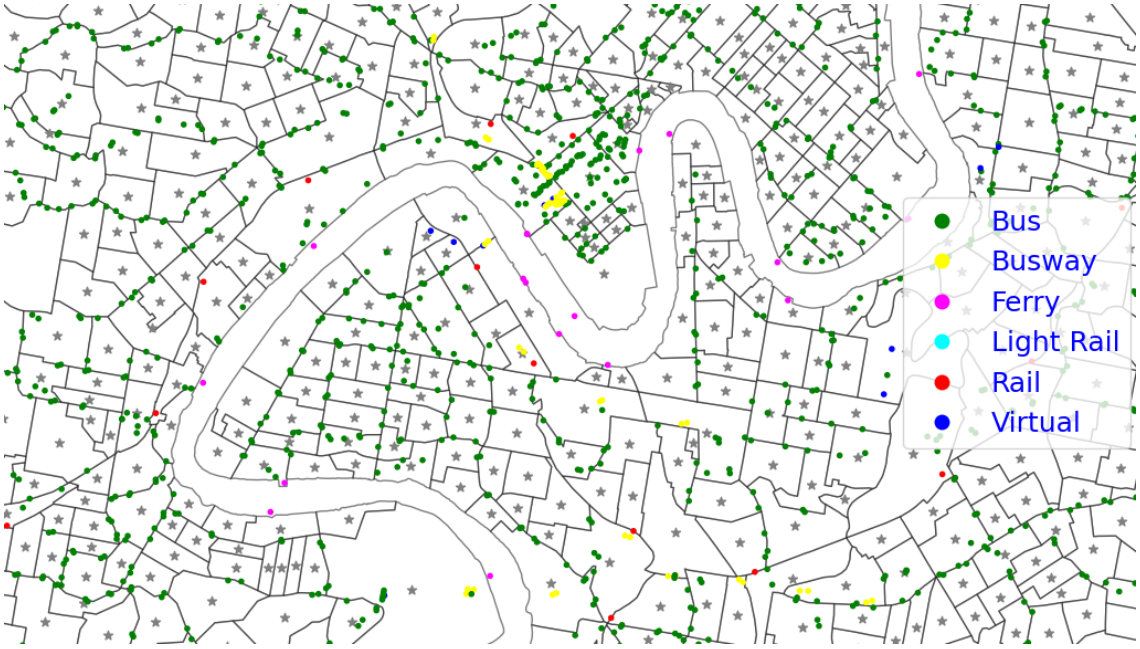

Figure 2.2: A set of PT stops and SA1 boundary along with centroids in grey is shown in different colours for a part of Queensland.

$$\sum_{i=1}^{n_o} p^O(i) = 1 \tag{2.1}$$

$$N_O \geq n_o \tag{2.2}$$

$$\frac{1}{N_O} \leq p^O(i) \leq \frac{n_o}{N_O} \tag{2.3}$$

Let, $p^D(j)$ be the probability of arriving at j^{th} PT stops in destination SA2.

$$\sum_{j=1}^{n_d} p^D(j) = 1 \tag{2.4}$$

$$N_D \geq n_d \tag{2.5}$$

$$\frac{1}{N_D} \leq p^D(j) \leq \frac{n_d}{N_D} \tag{2.6}$$

We know the number of trips between the origin PT stops i, and destination PT stops j either using the Markovian probability (for NSW ROAM data), or the number of trips (Queensland data). Let T_{ij} be the total trips between OD pair $OD(i, j)$. Out of these T_{ij} trips,

PT stop i is shared with no_1 number of SA1s. Out of those, no_2 SA1s belong to $SA2_O$. $\frac{no_2}{no_1}$ is the probability of PT stop i is boarded as an origin station from $SA2_O$.

PT stop j is shared with nd_1 number of SA1s. Out of those, nd_2 SA1s belong to $SA2_D$. $\frac{nd_2}{nd_1}$ is the probability of PT stop j is boarded as an origin station from $SA2_D$.

$$P(i,j) = T_{ij}\frac{nd_2}{nd_1}\frac{no_2}{no_1} \tag{2.7}$$

$$OD(SA2_O, SA2_D) = \sum_{i=1}^{n_o}\sum_{j=1}^{n_d} P(i,j)\cdot p^O(i)\cdot p^D(j) \tag{2.8}$$

2.4 SA1-based shared nearest stops

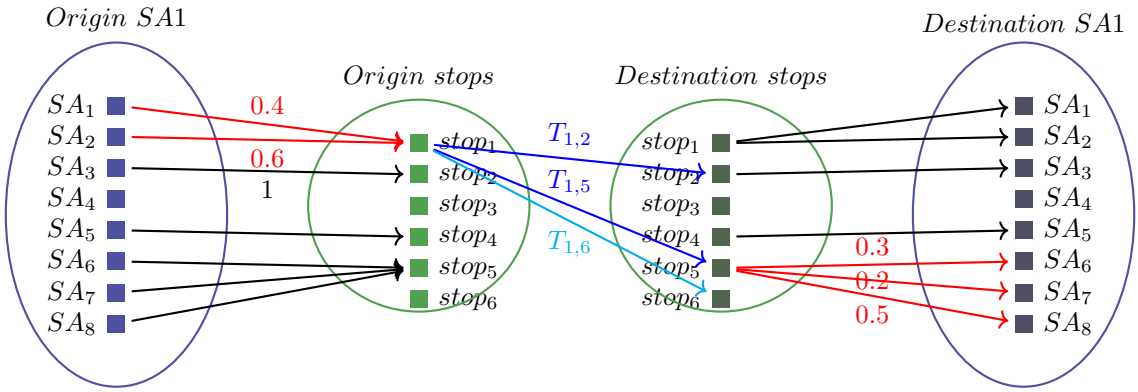

Figure 2.3: Mapping technique for PT stops to SA1 centroids based on SA1 based population probabilities.

2.4.1 Stop types

3 types of stops exist:

1. Stop1 and Stop5 are mapped to multiple SA1 as the closest PT stops. Distance-based probabilities are calculated.
2. Stop2 and Stop 4 are one-to-one mapped stops. Distance-based probabilities are 1.
3. Stop3 and Stop6 are not mapped to any SA1. But they are contained within any SA1 boundary.

$T_{1,2}$, $T_{1,5}$, $T_{1,6}$ are OD flows among stops, (i) resulting from the estimation model for NSW train data, and (ii) available for Queensland PT data. Let us consider, $SA3$ has two stops respectively, Stop2 and Stop3, such that Stop2 is the closest to the centroid. Similarly, $SA6$ has two stops, Stop5 and Stop6, within its geographic boundary.

Origin	Destination							
	$SA1$	$SA2$	$SA3$	$SA4$	$SA5$	$SA6$	$SA7$	$SA8$
$SA1$	0	0	$0.4 \cdot T_{1,2}$			$0.4 \cdot T_{1,5} \cdot 0.3$	$0.4 \cdot T_{1,5} \cdot 0.2$	$0.4 \cdot T_{1,5} \cdot 0.5$
$SA2$	0	0	$0.6 \cdot T_{1,2}$			$0.6 \cdot T_{1,5} \cdot 0.3$	$0.6 \cdot T_{1,5} \cdot 0.2$	$0.6 \cdot T_{1,5} \cdot 0.5$
$SA3$			0					
$SA4$				0				
$SA5$					0			
$SA6$						0		
$SA7$							0	
$SA8$								0

Table 2.1: Structure of SA1 based OD flow matrix derived from PT-based OD flow

1. All the stop-to-stop based OD flow is not considered. For example, $T_{1,6}$ is a valid OD flow from Stop1 to Stop6. But Stop6, although contained within SA6, is not the closest node to the centroid. Hence, $T_{1,6}$ is not included in the SA1-based OD flow calculation.
2. Stop3 is not closest to the centroid to any SA1, but there exists valid OD flow $T_{3,x}$, starting from Stop3, are not included. Similarly, all OD flow dissipating in Stop6 $T_{x,6}$ are not included.

	SA1_CODE21	SA2_CODE21	stop_id	stop_type	probability
4	30501110536	305011105	39	Bus	1
5	30501110506	305011105	49	Bus	0.51
6	30501110529	305011105	49	Bus	0.33
7	30501110533	305011105	49	Bus	0.165
8	30501110523	305011105	50	Bus	1
9	30501110542	305011105	52	Bus	1
10	30501110520	305011105	66	Bus	1
11	30501110519	305011105	72	Bus	1
12	30501110532	305011105	81	Bus	1
13	30501110541	305011105	90	Bus	1
14	30501110522	305011105	94	Bus	0.45
15	30501110526	305011105	94	Bus	0.55
16	30501110513	305011105	95	Bus	1

Table 2.2: Population-based probability calculation

2.5 Output OD flow data for QLD

Output OD matrices Q^k_{SA2} are square matrices with rows and columns as $SA2\_CODE21$. The cell values are the estimated number of movements happening for 4 calendar years, $2019-2022$. A number of OD Matrices are produced as output monthly and monthly-weekly basis. All the OD matrices

are stored in *.csv* format. A python script named $QLD\_SA2SA1\_PopulationtBasedProb\_ODmatrix.py$ is written to produce the SA2-based OD matrices for three modes of transport: (i) buses, (ii) rail (including trams and trains), and (iii) ferry.

2.5.1 Monthly-weekly OD matrices

There are monthly-weekly OD flow data available for 12 months for four years, separated by weekdays and weekends. For each month, weekdays and weekends OD flow data are separately used to generate the mobility among all SA2 pairs. Altogether, there are 288 OD flow matrices in either of the following format:

- $QLDRail\_Monthly\_OD\_Flow\_Matrix\_y\_m\_Weekday.csv$, such that m is the month and y is the year for all weekdays when the users availed rail as a mode of transport.
- $QLDRail\_Monthly\_OD\_Flow\_Matrix\_y\_m\_Weekend.csv$, such that m is the month and y is the year for all weekends when the users availed rail as a mode of transport.
- $QLDBus\_Monthly\_OD\_Flow\_Matrix\_y\_m\_Weekday.csv$, such that m is the month and y is the year for all weekdays when the users availed buses as a mode of transport.
- $QLDBus\_Monthly\_OD\_Flow\_Matrix\_y\_m\_Weekend.csv$, such that m is the month and y is the year for all weekends when the users availed busses as a mode of transport.
- $QLDFerry\_Monthly\_OD\_Flow\_Matrix\_y\_m\_Weekday.csv$, such that m is the month and y is the year for all weekdays when the users availed ferry as a mode of transport.
- $QLDFerry\_Monthly\_OD\_Flow\_Matrix\_y\_m\_Weekend.csv$, such that m is the month and y is the year for all weekends when the users availed ferry as a mode of transport.

2.5.2 Monthly OD matrices

There are 12 OD matrices for each year for each mode of transport. For example, $QLDBus\_Monthly\_OD\_Flow\_Matrix\_2019\_01.csv$ OD matrix contains the OD flow for the month of January for the year 2019 for all the passenger flow movement via buses between each possible pair of SA2s. Similarly, $QLDRail\_Monthly\_OD\_Flow\_Matrix\_2019\_12.csv$ and $QLDFerry\_Monthly\_OD\_Flow\_Matrix\_2019\_12.csv$ OD matrices respectively contain the OD flow for the month of December for the year 2019 for all the passenger flow movement via rails and ferries between each possible pair of SA2s.

Chapter 3

NSW SA2-level OD matrices

3.1 Introduction

This chapter contains the technical and operational metadata that is required to estimate origin-destination (OD) movement flow among the SA2 regions in NSW, connected by a train network. The Statistical Area 2 (SA2) regions of New South Wales connected by Sydney trains (T1-T9) and Metro services (Metro North West line) have been used to evaluate OD movement flows. The passenger OD movement data among different stations(or the station-based OD flow) are first estimated using a statistical estimation methodology. The stations-based OD flow data are then translated into region-based OD matrices using the state-of-art method. Below a brief description of the method is provided along with the input-output data, the assumptions and the future work that remains.

3.2 Method

In this section, a brief description of the method is described. A link count-based OD flow estimation model [Vardi, 1996, Lo et al., 1996, Yang et al., 1992, Dey et al., 2020] is implemented to produce the OD matrices of movement data. Centroids of each SA2 region have been considered as either origin or destination of a passenger who is travelling via the train network. Hence, there are 3 legs of a journey: (i) leg_1 travelling from SA2 to a railway station, (ii) leg_2 travelling from the origin station to a destination station, and (iii) leg_3 travelling from the destination station to the destination SA2 centroid. First, the station-based OD flow movement is estimated in the train network considering each pair of station nodes as an origin and destination. Next, each station OD pairs are mapped to each OD pair of SA2 centroids based on a set of assumptions according to the ABS census data.

3.2.1 Train station-based OD flow

Let G be a directed weighted train network graph, where the nodes represent the stations, and the edges represent links between two stations on a valid train route. Let there be n stations to have

$J = n\dot{(}n-1)$ possible OD pairs. Let I be the number of edges in G. Let ${Y_i}^k$ be the occupancy of passengers for a link i and ${X_j}^k$ be the number of passengers travelling between an OD pair j for a k^{th} day of a week, $k \in \{0, 1, 2, 3, 4, 5, 6\}$ such that $k = 0$ for Monday and so on. Let $\mathbf{Y}^k$ represent the set of link-wise occupancy data and $\mathbf{X}^k$ the OD flow vector such that:

$$\begin{aligned} \mathbf{X}^k &= ({X_1}^k, \cdots, {X_j}^k, \cdots, {X_J}^k)^\top \\ \mathbf{Y}^k &= ({Y_1}^k, \cdots, {Y_i}^k, \cdots, {Y_I}^k)^\top \end{aligned} \tag{3.1}$$

Let p_{ij} be the route choice probabilities defined as the probability of a passenger travelling with an OD pair j getting counted in the link i. We obtain the following route choice probability matrix P:

$$P = \begin{matrix} \\ \text{Links } (I) \end{matrix} \begin{matrix} \text{OD pairs } (J) \\ \begin{bmatrix} p_{11} & p_{12} & \cdots & p_{1J} \\ p_{21} & p_{22} & \cdots & p_{2J} \\ \vdots & \vdots & \ddots & \vdots \\ p_{I1} & p_{I2} & \cdots & p_{IJ} \end{bmatrix} \end{matrix} \tag{3.2}$$

P is constructed according to a markovian routing assumption of the transport network tomography model [Vardi, 1996, Dey et al., 2020] from the weighted graph G, and the following is obtained:

$$E[\mathbf{Y}^k] = \begin{bmatrix} \mathrm{E}[Y_1] \\ \mathrm{E}[Y_2] \\ \vdots \\ \mathrm{E}[Y_I] \end{bmatrix} = \mathbf{P}\mathbf{X}^\mathbf{k} \tag{3.3}$$

$E[\mathbf{Y}^k]$ is obtained from the link occupancy data. Then, a Landweber iteration [Landweber, 1951] is used to solve Equation 3.3. A solution of Equation 3.3 is estimated using a linear least squares gradient. Given an $I \times J$ matrix P and vector Y, our goal is to find a vector $X \in \mathbb{R}^n$ such that the following objective function is minimized with a predefined bounds $\mathbf{X^U} \in \mathbb{R}^n$ and $\mathbf{X^L} \in \mathbb{R}^n$:

$$\begin{aligned} f(X) &= \frac{1}{2I}\sum_{i=1}^{I}({\mathbf{p_i}}^\top X - Y_i)^2 = \frac{1}{2I}\|PX - Y\|^2 \\ \nabla f(X) &= P^\top(PX - Y) \\ \nabla^2 f(X) &= P^\top P \\ \mathbf{X^L} &\geq \mathbf{X} \geq \mathbf{X^U} \end{aligned} \tag{3.4}$$

The solution to the Equation 3.3 provides us the estimated passenger flow for each OD pair j, which then is translated to SA2 level mobility.

3.2.2 SA2 based OD flow using distance-based probability assuption

$SA2_{cen} = \left\{C^1, \cdots, C^N\right\}$ be the set of all N centroids of the SA2 regions. There are $S = N\dot{(}N-1)$ OD pairs. Let sG be a directed weighted graph such that each node in $SA2_{cen}$ is connected to all the n nodes of G. The weights of sG are the euclidean distances between the pair of nodes. For

an SA2 OD pair $\{C^p, C^q\} = s_{pq} \in \{1, \cdots, S\}$, let the origin station and destination station of leg_2 belong to an OD pair $j(s_pq) \in \{1, \cdots, J\}$ of G. Then, for k^{th} day of the week, the estimated OD flow X^k_{js} is mapped to the OD pair $s^k_{pq} = \{C^p, C^q\}$. The output OD matrix M^k_{SA2} is defined as follows:

$$M^k_{SA2} = \begin{array}{c} \\ \text{Origins } (N) \end{array} \begin{array}{c} \text{Destinations } (N) \\ \begin{bmatrix} s^k_{11} & s^k_{12} & \cdots & s^k_{1N} \\ s^k_{21} & s^k_{22} & \cdots & s^k_{2N} \\ \vdots & \vdots & s^k_{pq} & \vdots \\ s^k_{N1} & s^k_{N2} & \cdots & s^k_{NN} \end{bmatrix} \end{array} \tag{3.5}$$

$$s^k_{pq} = X^k_{j(s_{pq})}$$

$$diag\{M^k_{SA2}\} = s^k_{11} = s^k_{22} = \cdots = s^k_{NN} = 0$$

3.2.3 SA2 based OD flow using population-based probability assumption

This assumption is similar to the method described in section 2.3.1, where the mapping of station-based to SA2-based OD flow is converted using a population-based probability. Once the station-based OD flow is estimated, the next step is to assign the probability of population mobility of an SA2 commuting via each origin station considering the destination station is assigned with a probability with the destination SA2. Figure 3.1 is illustrating a set of the NSW train stations with red triangular markers with SA2 boundaries in black and SA1 boundaries in blue. It is evident that many SA1s within the SA2s do not contain any nearest station, but can be assigned with a probability to access an origin station or destination after reaching a destination station, according to assumptions stated in section 2.3.1. The final output is generated depending upon the assumption of modelling choice.

3.3 Input data

All the input data for the model is preprocessed from the raw roam data files. The raw data are available from ROAM (Rail Opal Assignment Model) dataset [Transport for NSW, 2022]. This dataset contains information about passenger occupancy on trains. Occupancy is recorded at different times and stations for each day. Passenger counts are grouped into ranges of 20 and are based on arrivals at the stop. The data set has a spatial coverage of Sydney trains (T1-T9) and Metro services (Metro North West line). Then, a set of python based scripts is developed to process the raw roam data to model input followed by a set of intermediate data files. This section provides a description of the codes and the process of generating the input data files and their structure.

3.3.1 Edge-list occupancy data

Raw roam data files are available in *.txt* format for each day, e.g. $ROAM\_20190904.txt$ is a file generated on September 4, 2019. The raw file contains 24 columns including a current stop as ACT_Stop_STN, type of card as $CARD\_TYPE$ and a probable range of passenger occupancy as $OCCUPANCY\_RANGE$. The information is translated into cumulative occupancy for all card

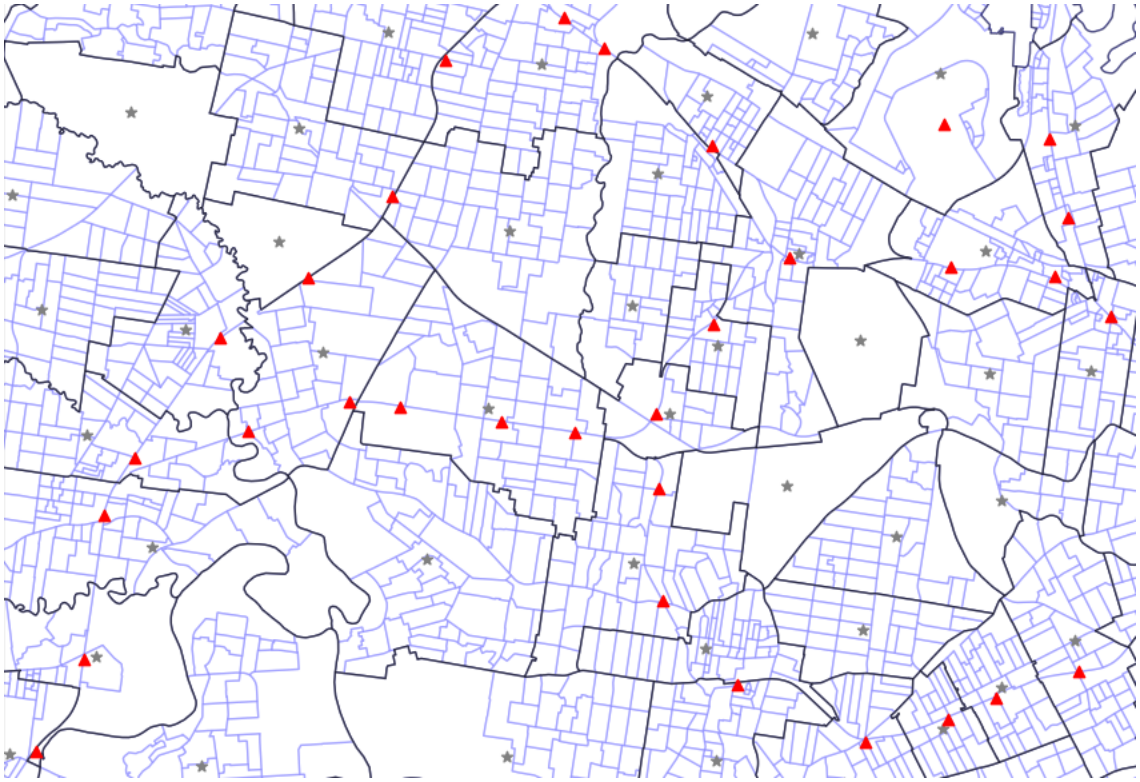

Figure 3.1: Part of the SA1 boundary in blue and SA2 centroids in grey train stops are shown in red colour.

types and an edge-list table is constructed with the name *ROAM_LinkCount.csv*, which contains 122944 number of row elements from the date 01/07/2019 to 30/11/2019. The first 10 elements of the *ROAM_LinkCount.csv* are shown in Table 3.1. The links are constructed as a set of present stop and next stop from a set of data frames, grouped by the columns $SEGMENT\_DIRECTION$, $TRIP\_NAME$, $SERVICE\_LINE$ of the raw roam data files. Then, the occupancy data along with data and day-of-week (dow) information are stacked into one single *.csv* file from all the raw roam files using a Python script *ROAM_Create_linkCountTable.py*.

3.3.2 The networks G and sG

The network G and sG are constructed as a directed and weighted graph using a python script named *Roam_Create_Graphs.py*. Input to the codes is a set of preprocessed files named as *nsw_stations_with_nodes.csv*, *nsw_sa2s_with_centroids.csv*, *nsw_path_to_station.csv* and *ROAM_LinkCount.csv*. The output of the code is two directed weighted graphs G, sG and one CSV file *OD_main_Sim1.csv* containing the information of the mapping SA2 centroids OD pair to station OD pairs.

G is constructed from the *ROAM_LinkCount.csv* file utilising the $networkx.from\_pandas\_edgelist()$

PresentStop	NextStop	Occupancy	Date	dow	Link
Aberdeen	Muswellbrook	20	30/11/2019	5	('Aberdeen', 'Muswellbrook')
Aberdeen	Scone	20	30/11/2019	5	('Aberdeen', 'Scone')
Adamstown	Broadmeadow	1180	30/11/2019	5	('Adamstown', 'Broadmeadow')
Adamstown	Kotara	950	30/11/2019	5	('Adamstown', 'Kotara')
Albion Park	Dapto	890	30/11/2019	5	('Albion Park', 'Dapto')
Albion Park	Oak Flats	950	30/11/2019	5	('Albion Park', 'Oak Flats')
Allawah	Carlton	5760	30/11/2019	5	('Allawah', 'Carlton')
Allawah	Hurstville	7450	30/11/2019	5	('Allawah', 'Hurstville')
Arncliffe	Banksia	11070	30/11/2019	5	('Arncliffe', 'Banksia')
Arncliffe	Wolli Creek	9660	30/11/2019	5	('Arncliffe', 'Wolli Creek')

Table 3.1: An instance of first 10 elements of the edge-list occupancy data table *ROAM_LinkCount.csv*.

function. Weights of the Graph G are calculated using a file named *nsw_stations_with_nodes.csv*, containing the stop names geometry. The geometry column has the latitude, and longitude of the stops. The geometry column is converted to a projected CRS and The euclidian distances between each pair of stops are calculated. An illustration of the train stops with their projected geometry is presented in Figure 3.2.

Figure 3.3 is presenting part of the Sydney metro rail network along with the SA2 centroids. The SA2 centroids are shown with blue dots and the station names are shown in green circles. Black arrows are drawn between the green circles to demonstrate the edges/links in the train network G. The Train network sG is constructed connecting each SA2 centroid to all the station nodes of G. The network sG is used to find the leg_1, leg_2, and leg_3 distances for each SA2 centroid pair using the shortest path assumption. All the shortest paths and the lengths of the shortest paths in sG are calculated using the $networkx.floyd\_warshall\_predecessor\_and\_distance()$ function. The mapping $s^k_{pq} = X^k_{j(s_{pq})}$ is provided between the columns *origin* and *destination* to O_{leg2} and D_{leg2} respectively in the file *OD_main_Sim1.csv*. The *origin* and *destination* columns are containing the SA2 centroids, and the O_{leg2} and D_{leg2} columns are containing the mapped stations for the train journey part.

Assumptions

The mapping has been done based on the following assumptions from the ABS census data:

1. A person starts the journey from the origin SA2 centroid and follows the shortest path in the graph sG to its destination centroid such that leg_2 contains at least two station nodes.
2. No person will travel more than 250 km daily.
3. Distance travelled in leg_2 is always greater than the distances travelled in other legs i.e. $leg_2 > leg_1$ and $leg_2 > leg_3$.
4. There is no transfer happening during the leg2 journey i.e. leg_2 does not contain any intermediate SA2 centroids.

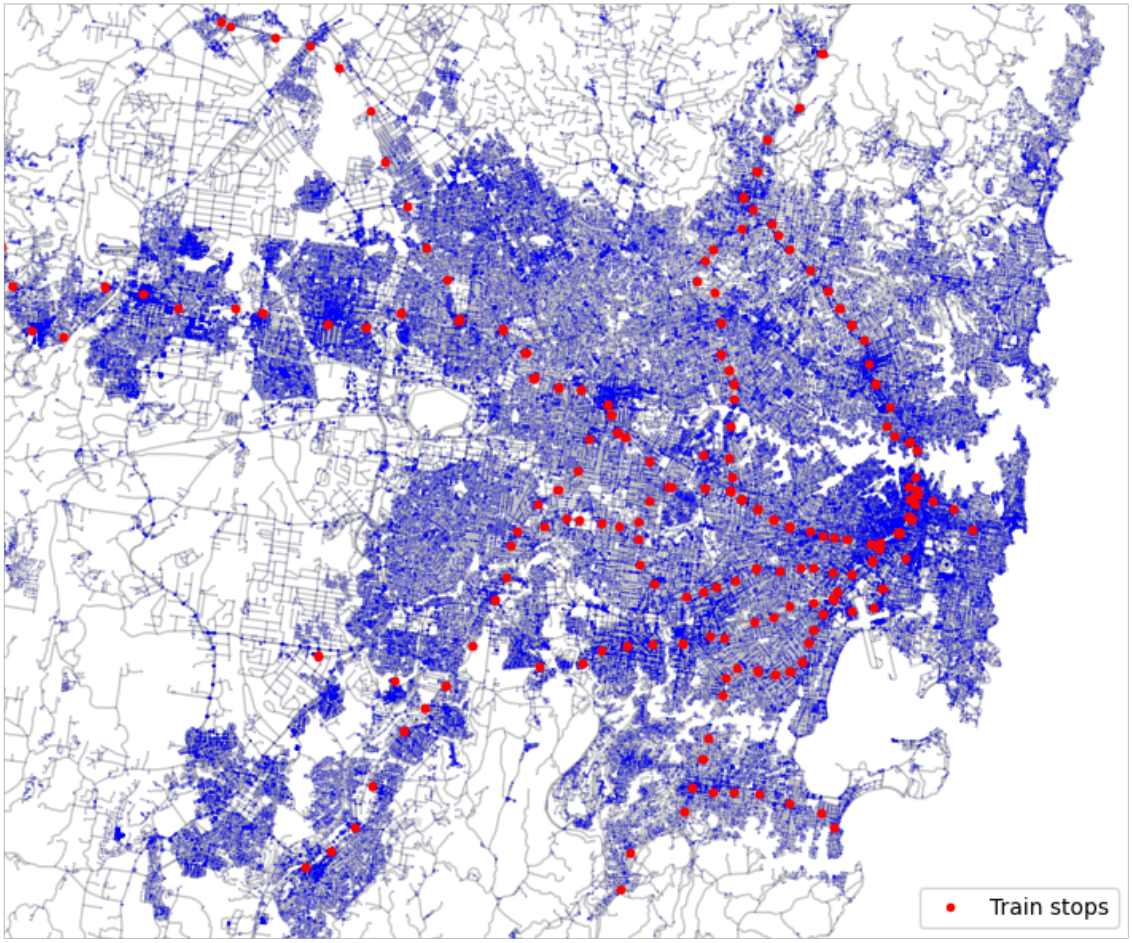

Figure 3.2: A set of train stops is shown in red colour for a part of the Sydney rail network.

3.3.3 P data

The Markovian routing probability matrix P in Equation 3.2 is constructed using a python script *Roam_Create_P_OD.py*. The 3 input files for the code *Roam_Create_P_OD.py* are *ROAM_LinkCount.csv*, *ODColdf.csv*, and *nsw_stations_with_nodes.csv*. The P matrix is provided as the output file name *Roam_OD_prob_M.csv*.

3.3.4 $\mathbf{Y}^k$ data

$\mathbf{Y}^k$ data is generated from the table *ROAM_LinkCount.csv* providing the value of $k = dow \in \{0, 1, 2, 3, 4, 5, 6\}$. For example, for each Wednesday, we generate the data frame $\mathbf{Y}^2$ as $k = 2$ and the data frame is stored as *Roam_Y_2019_dow_2.csv*. Table 3.2 is presenting the weekly Wednesdays in the columns and occupancy data of the links in the rows. $\{E[Y_1], \cdots, E[Y_I]\}$ are constructed

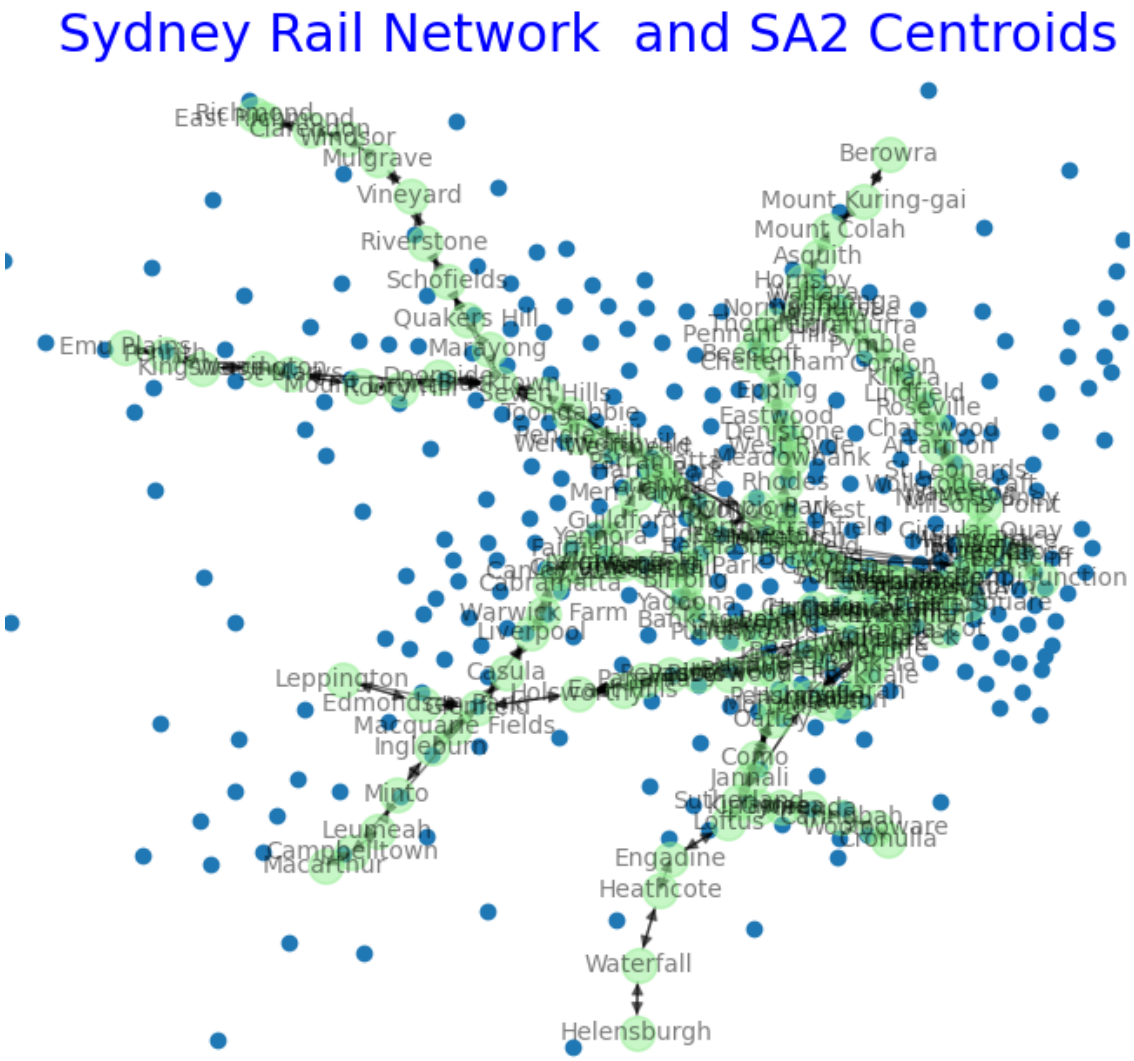

Figure 3.3: A set of SA2 centroids (in blue dots) are illustrated with a part of the Sydney metro rail network part.

taking the row-wise average for each link in Table 3.2.

3.3.5 Estimation for $\mathbf{X}^k$ data

A python script named *ROAM_EM_estimattion.py* is written that generates the $E[\mathbf{Y}^k]$ from the file *ROAM_LinkCount.csv*, with the value of k is provided as the variable name *dow* within the script. Then the *Roam_OD_prob_M.csv* is used as the P matrix. The file *ODColdf.csv* is provided as input to provide the $\mathbf{X^U}$ and $\mathbf{X^U}$. Finally, Equation 3.4 is solved using an iterative method of

Links	2019-11-27	2019-11-20	2019-11-13	2019-06-11	2019-10-30
(Aberdeen, Muswellbrook)	90	0	30	30	30
(Aberdeen, Scone)	90	0	30	30	30
(Muswellbrook, Aberdeen)	90	0	30	30	30
(Muswellbrook, Singleton)	120	0	40	30	40
(Scone, Aberdeen)	90	0	30	30	30
(Adamstown, Broadmeadow)	1540	1320	1220	1120	1340
(Adamstown, Kotara)	1050	790	770	750	880
(Broadmeadow, Adamstown)	890	690	610	650	640
(Broadmeadow, Hamilton)	2880	2420	2160	2040	2320
(Broadmeadow, Cardiff)	1350	950	930	910	830

Table 3.2: An instance of first 10 elements of $\mathbf{Y}^2$ stored in the table *Roam_Y_*2019*_dow_*2*.csv*

gradient descent as suggested by Landweber iteration [Landweber, 1951].

3.3.6 Estimation results

There are 7 generated files that contain station-based OD flow for each day of the week within the temporal coverage from July 2019 to November 2019:

1. *Roam_Estimated_OD_*2019*_dow_*0*.csv* file contains the OD flow on Mondays.
2. *Roam_Estimated_OD_*2019*_dow_*1*.csv* file contains the OD flow on Tuesdays.
3. *Roam_Estimated_OD_*2019*_dow_*2*.csv* file contains the OD flow on Wednesdays.
4. *Roam_Estimated_OD_*2019*_dow_*3*.csv* file contains the OD flow on Thursdays.
5. *Roam_Estimated_OD_*2019*_dow_*4*.csv* file contains the OD flow on Fridays.
6. *Roam_Estimated_OD_*2019*_dow_*5*.csv* file contains the OD flow on Saturdays.
7. *Roam_Estimated_OD_*2019*_dow_*6*.csv* file contains the OD flow on Sundays.

3.4 Output OD flow data for NSW based on distance-based assumptions

Output OD matrices M^k_{SA2} are square matrices with rows and columns as $SA2\_CODE21$. The cell values are the estimated number of movements happening within the temporal coverage. The temporal coverage for the output is from July 2019 to November 2019, and for the year 2020. A number of OD Matrices are produced as output weekly and monthly basis. All the OD matrices are stored in *.csv* format. A python script named $ROAM\_EM\_estimateToODflow.py$ is written to convert the estimated flow, i.e. the 7 files generated in the previous section, to SA2-based OD matrices.

3.4.1 Yearly OD matrices

There are 7 OD matrices for each day of the week within the temporal coverage from July 2019 to November 2019:

1. $NSWTrains\_Yearly\_OD\_flow\_Matrix\_2019\_dow\_0.csv$ matrix contains the OD flow on Mondays.
2. $NSWTrains\_Yearly\_OD\_flow\_Matrix\_2019\_dow\_1.csv$ matrix contains the OD flow on Tuesdays.
3. $NSWTrains\_Yearly\_OD\_flow\_Matrix\_2019\_dow\_2.csv$ matrix contains the OD flow on Wednesdays.
4. $NSWTrains\_Yearly\_OD\_flow\_Matrix\_2019\_dow\_3.csv$ matrix contains the OD flow on Thursdays.
5. $NSWTrains\_Yearly\_OD\_flow\_Matrix\_2019\_dow\_4.csv$ matrix contains the OD flow on Fridays.
6. $NSWTrains\_Yearly\_OD\_flow\_Matrix\_2019\_dow\_5.csv$ matrix contains the OD flow on Saturdays.
7. $NSWTrains\_Yearly\_OD\_flow\_Matrix\_2019\_dow\_6.csv$ matrix contains the OD flow on Sundays.

3.4.2 Monthly-weekly OD matrices

There are monthly-weekly OD flow data available for 5 months of 2019: *July*, *August*, *Septemer*, *October*, *November* and *December*. Each month has 7 OD flow matrices for each day of the week. Altogether, there are 35 OD flow matrices in the file format:
$NSWTrains\_Monthly\_OD\_flow\_Matrix\_2019\_Month\_m\_dow\_w.csv$, such that x is the month and w is the day of the week. For example, $Monthly\_OD\_flow\_Matrix\_Month\_07\_dow\_1.csv$ contains the OD flow on each Tuesday for the month of July 2019, and so on.

3.5 Output OD flow data for NSW based on population-based assumptions

Similarly, a set of output OD matrices M^k_{SA2} are square matrices with rows and columns as $SA2\_CODE21$ generated based on population based assuptions. The only difference is with the python script, $NSW\_ROAM\_SA2SA1\_PopulationtBasedProb\_ODmatrix.py$ is created to produce the monthly and weekly OD flow matrices. The python script needs the input files named $all\_australia\_sa1\_population\_and\_centroids\_with\_sa2\_commuting.csv$ that contains the details of SA1 based population and commutating statistics, and the location of the station coordinates located in file $nsw\_stations\_with\_nodes.csv$. The python script also needs the shape file, named as $SA1\_2021\_AUST\_GDA2020.shp$ to produce the desirbale OD flow output files.

3.6 Comments

There are issues that we are currently working to solve.

1. There are two solving techniques used for estimating the station-based OD flow. The Landweber iteration is used to solve the Linear inverse ill-posed problem. A bounded search on a sparse matrix space is done by a Projected Gradient Descent implementation in Python. This part has been reproduced with a solver to speed up the simulation and achieve better estimation accuracy. For this work, Gurobipy is used with the same upper and lower bounds described earlier.
2. In the gradient descent algorithm, the Landweber iteration suggests an upper limit of the step size which is needed to be calculated after the Singular Value decomposition (SVD) of the P matrix. The highest singular value is used to calculate the Landweber iteration suggested step size. We are unable to calculate the step size as of now due to the memory allocation error. Hence the step size is currently variable between set 1 for 150 iteration. These values will be changed once the SVD of P can be found.

Bibliography

Hugo Barbosa, Marc Barthelemy, Gourab Ghoshal, Charlotte R James, Maxime Lenormand, Thomas Louail, Ronaldo Menezes, José J Ramasco, Filippo Simini, and Marcello Tomasini. Human mobility: Models and applications. *Physics Reports*, 734:1–74, 2018.

Subhrasankha Dey, Stephan Winter, and Martin Tomko. Origin–destination flow estimation from link count data only. *Sensors*, 20(18):5226, 2020.

Marta C Gonzalez, Cesar A Hidalgo, and Albert-Laszlo Barabasi. Understanding individual human mobility patterns. *nature*, 453(7196):779–782, 2008.

Louis Landweber. An iteration formula for fredholm integral equations of the first kind. *American journal of mathematics*, 73(3):615–624, 1951.

HP Lo, N Zhang, and William HK Lam. Estimation of an origin-destination matrix with random link choice proportions: A statistical approach. *Transportation Research Part B: Methodological*, 30(4):309–324, 1996.

Dunstan Matekenya, Xavier Espinet Alegre, Fatima Arroyo Arroyo, and Marta Gonzalez. Using mobile data to understand urban mobility patterns in freetown. *Sierra Leone*, 2021.

Mohamed Mokbel, Mahmoud Sakr, Li Xiong, Andreas Züfle, Jussara Almeida, Walid Aref, Gennady Andrienko, Natalia Andrienko, Yang Cao, Sanjay Chawla, et al. Towards mobility data science (vision paper). *arXiv preprint arXiv:2307.05717*, 2023.

Xuan Song, Hiroshi Kanasugi, and Ryosuke Shibasaki. Deeptransport: Prediction and simulation of human mobility and transportation mode at a citywide level. In *Proceedings of the twenty-fifth international joint conference on artificial intelligence*, pages 2618–2624, 2016.

Transport for NSW. ROAM - Rail Opal Assignment Model. `https://opendata.transport.nsw.gov.au/dataset/roam-rail-opal-assignment-model`, 2022.

Y Vardi. Network Tomography : Estimating Source-Destination Traffic Intensities From Link Data. *Journal of the American Statistical Association*, 1459(April 2013):37–41, 1996.

Yanyan Xu, Luis E Olmos, David Mateo, Alberto Hernando, Xiaokang Yang, and Marta C Gonzalez. Urban dynamics through the lens of human mobility. *arXiv preprint arXiv:2305.16996*, 2023.

Hai Yang, Tsuna Sasaki, Yasunori Iida, and Yasuo Asakura. Estimation of origin-destination matrices from link traffic counts on congested networks. *Transportation Research Part B: Methodological*, 26(6):417–434, 1992.