Leveraging GTFS data to assess transit supply

James Reynolds[a,1,*], Yanda Qu[a,2], Graham Currie[a,3]

[a] *Public Transport Research Group (PTRG), Institute of Transport Studies, Department of Civil Engineering Engineering, Monash University, Clayton Campus, Melbourne, 3800, Victoria, Australia*

Abstract

This is the abstract.

It consists of two paragraphs.

Keywords: keyword1, keyword2

1. Introduction

“If you can’t measure it, you can’t manage it” is often miss-attributed to Deming (1993) who, according to Berenson (2016), was actually trying to make the opposite point. Regardless, service level indicators are important in researching, managing and seeking to improve transit operations (Fielding, 1987; Ryus et al., 2003). Many indicators already exist including, for example: those in the Transit Capacity and Quality of Service Manual (TCQSM)(Kittleson & Associates et al., 2013); and the Transit Score metric (Walk Score, 2023). There are two inter-related challenges in using such metrics: (1) calculating the scores themselves for a specific location and service pattern; and (2) understanding the metrics, their meaning and importance (and being able to explain this to those who are not specialists in transit). For example: the TCQSM metrics might fail the first challenge, being difficult to calculate without specialist software and data. However, they use an A to F scoring system and there is an entire guidebook explaining each individual indicator, suggesting that they might pass the second. In contrast, entering an address into the Transit Score website will return a score out of 100 reflecting the quantity of transit available, which is both easy to obtain and explain. However, the methodology and algorithm behind the Transit Score is not publicly available, so these cannot be calculated independently. This might limit practitioners, researchers, advocates or others involved in transit planning, operations or policy-making from reporting score changes associated with new infrastructure, service patterns or improvement options.

Previous research by Currie and Senbergs (2007) developed a transit Supply Index (SI), This appears relatively easy to understand and explain, as it is based on the number of transit arrivals at stops within an area of interest.

An adjustment is made to account for how much of the area of interest is within walking distance of each stop, meaning that higher SI scores indicate areas with more frequent services and/or better coverage.

Unfortunately, the SI does not appear to have been widely used, perhaps in part because at the time it was first published it was not particulrly easy to calculate. At that time timetable data was not publicly available in a standardized and machine-readable format, and the scores reported in Currie and Senbergs (2007) were calculated directly from a database of services provided by the transit authority in Melbourne,

[*] Corresponding author

Email addresses: james.reynolds@monash.edu (James Reynolds), yanda.qu@monash.edu (Yanda Qu), graham.currie@monash.edu (Graham Currie)

[1] Research Fellow

[2] PhD Strudent

[3] Professor

Australia. Nowadays, the General Transit Feed Specification (GTFS) allows timetable publication in a standardized format, with more than 10,000 agencies providing feeds[4] (MobilityData, undated),

Many visualization, processing and analysis tools that accept GTFS data are now available. A gap, however, is that there is not yet a tool to calculate SI scores directly from GTFS datasets. This provides the motivation for the research reported in this paper, in which a new R package (gtfssupplyindex) specifically developed to calculate SI scores is presented. The remainder of this paper is structured as follows: the next section outlines the background to this research, including the original formulation of the Transit Supply Index, and an explanation of the GTFS. Section 3 then describes the study methodology, followed by a brief presentation of results in Section 4. Section 5 discusses the results, outlines directions for future research and provides a brief conclusion.

2. Background

2.1. Transit metrics

Even a brief search reveals many metrics available for benchmarking transit services. Examples include: (1) those in the Transit Cooperative Research Program (TCRP) Report 88, which is an extensive guidebook on developing a performance-measurement system (Ryus et al., 2003); (2) online databases provided by the Florida Transit Information System (FTIS) (Florida Transit Information System, 2018) and International Association of Public Transport (UITP) (2015); (3) those used in the extensive annual benchmarking program undertaken yearly by the Transport Strategy Centre in the United Kingdom, including over 100 transit providers around the world (Imperial College London, undated); and (4) a recently developed methodology to calculate ‘blank spots’ within an area, being those places beyond 400/800 metre walking distances to/from bus and tram stops/train stations (Alamri et al., 2023).

The Fielding Triangle (Fielding, 1987) provides a framework for understanding how such metrics combine service inputs, outputs and consumption to describe cost efficiency and effectiveness; and service effectiveness. More broadly Litman (2003) and Litman (2016) discuss some of the traffic, mobility, accessibility, social equity, strategic planning and other rational decision-making-based perspectives underling such metrics, while Reynolds et al. (2017) extends these into models of how institutionalism, incrementalism and other public policy analysis concepts might apply to decision-making processes relating to transit prioritisation. Guzman et al. (2017), developed a measure of accessibility in the context of policy development and social equity for Latin American Bus Rapid Transit (BRT) networks, while Creutzig et al. (2020) introduced street space allocation metrics based around 10 ethical principles

However, many of these metrics appear difficult to calculate, complex to explain or understand, and likely not well suited to communication with those who are not transit planners or engineers, or other technical specialists. Where pre-calculated metrics are immediately available it may not be possible for practitioners, researchers or advocates to independently generate metrics for proposed system changes. Sometimes it is not even possible to know precisely how scores for the existing services levels are calculated. For example, Transit Scores for locations with a published GTFS feed are readily available on the Walk Score (2023) website, eliminating the need for any calculations. The meaning of these Transit Scores appears easy to explain, as the highest possible score of 100 represents what might be experienced in the centre of New York. However, the Transit Score algorithm is patented and effectively a black box. It is not possible to calculate Transit Scores scores independently. Nor can Transit Scores to be generated for proposed changes to networks. The Transit Score metric, therefore, fails the first of the aforementioned challenges, as practitioners, researchers and advocates can only use those scores provided online. But, because it is based on a patented algorithm it may not be easy to understand or explain the connection between real-world conditions and the Transit Score, or what might need to be done to improve the score (and service levels). As such, it might partially pass the second of the aforementioned challenges, as the score’s concept is simple-
the closer to 100, the better - but further detail is limited.

[4] There are two forms: GTFS-static consisting of the timetable data (the scheduled services); and GTFS-realtime, which includes vehicle arrivals and departure times based on real-world position data. This paper and project uses only the GTFS-static (timetable) format.

In contrast, the TCQSM, specifies Levels of Service (LOS) between A and F across a range of factors[5]. This scoring scheme appears relatively simple to explain - A is good and F is bad - and matches that often used in traffic capacity analysis. Extensive detail is provided within Kittleson & Associates et al. (2013) scores mean. However, calculation of many of TCQSM metrics may need specialised software and datasets[6] and it might be challenging to explain the detail of these measures or how to improve them to non-technical decision-makers, stakeholders or others involved in transit management or advocacy.

2.2. GTFS

The General Transit Feed Specification (GTFS) is an open, text-based format that was developed originally to allow transit information to be included in the Google Maps navigation platform (MobilityData, undated). Figure @ref(fig:GTFS_ERD) shows an Entity Relationship Diargram (ERD) of the GTFS data structure, in which each box represents a database table in the GTFS. Table rows indicate the variables (columns) included in each, for example
each record in the 'stops' table includes a value for stop_id, stop_name, stop_lat and stop_lon. Relationships between the tables are indicated by the connecting lines, and Primary Key (PK) and Foreign Key (FK) designations, for example, stop_id appears in the 'stops' and 'stop_times' tables as a Primary Key and Foreign Key. 'Crow's feet' indicate the relationships between each table[7]

GTFS allows individual transit systems to be included in many online products and analysis, including the Transit Score metric itself. Wong (2013) provides another example of what can be done with GTFS data, having developed code to calculate of some of the TCQSM metrics and compared these across 50 transit operators. The Wong (2013) code is readily available (https://github.com/jcwong86/GTFS_Explore_Tool), but does not appear to be currently maintained. Future research may involve reviewing this code and using it to analyse modern GTFS feeds,
but in this paper the aim is more modest, being to use GTFS data to calculate Currie and Senbergs' (2007) SI.

2.3. The Transit Suppy Index

Currie and Senbergs' (2007) focus was the context of Melbourne's Census Collection Districts (CCD) and calculations based on a week of transit service. A more generalized form of the Transit Supply Index (SI) is shown below:

$$SI_{area,time} = \sum \frac{Area_{Bn}}{Area_{area}} * SL_{n,time}$$

(1) $SI_{area,time}$ is the Supply Index for the area of interest and a given period of time;
(2) $Area_{Bn}$ is the buffer area for each stop (n) within the area of interest. In Currie and Senbergs (2007) this was based on a radius of 400 metres for bus and tram stops, and 800 metres for railway stations;
(3) $Area_{area}$ is the area of the area of interest; and
(4) $SL_{n,time}$ is the number of transit arrivals for each stop for a given time period.

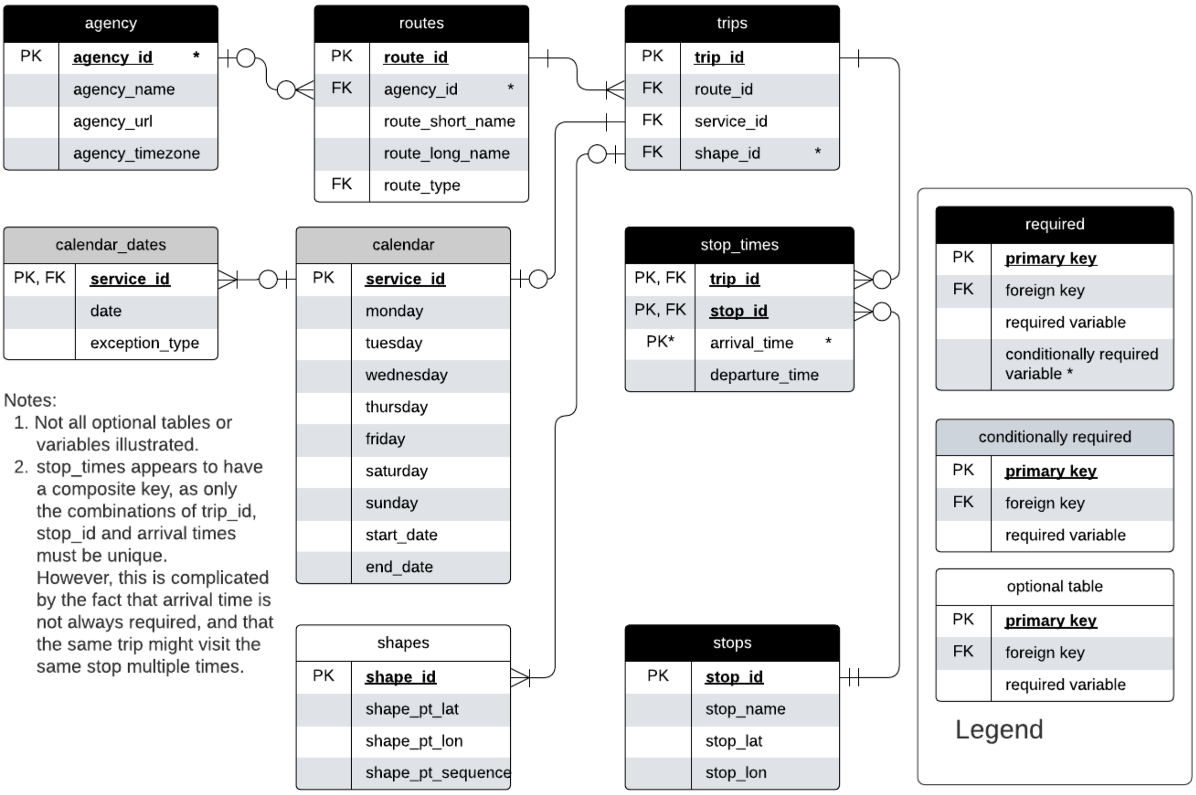
An advantage of the SI is that it is a relatively simple number to calculate, understand and explain. It describes the number of transit arrivals at stops within an area of interest and time frame, multiplied by a factor accounting for the proportion of the area of interest that is within typical walking distances of each stop. Hence, more services, more stops and higher frequencies increase the score.

The SI score does not incorporate service span, speed or other elements of a transit service. While these can be important to passenger experience, they might add complexity. Simplicity is also helped by the way that the SI is additive, in that $SI_{area,time}$ scores can be aggregated to calculate an overall score across multiple time periods or for a region encompassing multiple areas of interest.

[5] Including service span, frequency, speed, the proportion of the population serviced, competitiveness of travel times to car-based travel, and many more.

[6] For example, the Service Coverage Area metric in the TCQSM (pp. 5-8 to 5-21) may require GIS or other analysis, on top of accurate data about population densities, stop locations and service schedules.

[7] See https://i.stack.imgur.com/fxaAq.png for guide to the symbols.



agency	
PK	**agency_id** *
	agency_name
	agency_url
	agency_timezone

routes	
PK	**route_id**
FK	agency_id *
	route_short_name
	route_long_name
FK	route_type

trips	
PK	**trip_id**
FK	route_id
FK	service_id
FK	shape_id *

calendar_dates	
PK, FK	**service_id**
	date
	exception_type

calendar	
PK	**service_id**
	monday
	tuesday
	wednesday
	thursday
	friday
	saturday
	sunday
	start_date
	end_date

stop_times	
PK, FK	**trip_id**
PK, FK	**stop_id**
PK*	arrival_time *
	departure_time

shapes	
PK	**shape_id**
	shape_pt_lat
	shape_pt_lon
	shape_pt_sequence

stops	
PK	**stop_id**
	stop_name
	stop_lat
	stop_lon

Notes:
1. Not all optional tables or variables illustrated.
2. stop_times appears to have a composite key, as only the combinations of trip_id, stop_id and arrival times must be unique. However, this is complicated by the fact that arrival time is not always required, and that the same trip might visit the same stop multiple times.

required	
PK	**primary key**
FK	foreign key
	required variable
	conditionally required variable *

conditionally required	
PK	**primary key**
FK	foreign key
	required variable

optional table	
PK	**primary key**
FK	foreign key
	required variable

Legend

Figure 1: GTFS entity relationship diagram. Source: adapted by author from Alamri et al (2023) and the GTFS Schedule Reference (16/11/2023 revision).

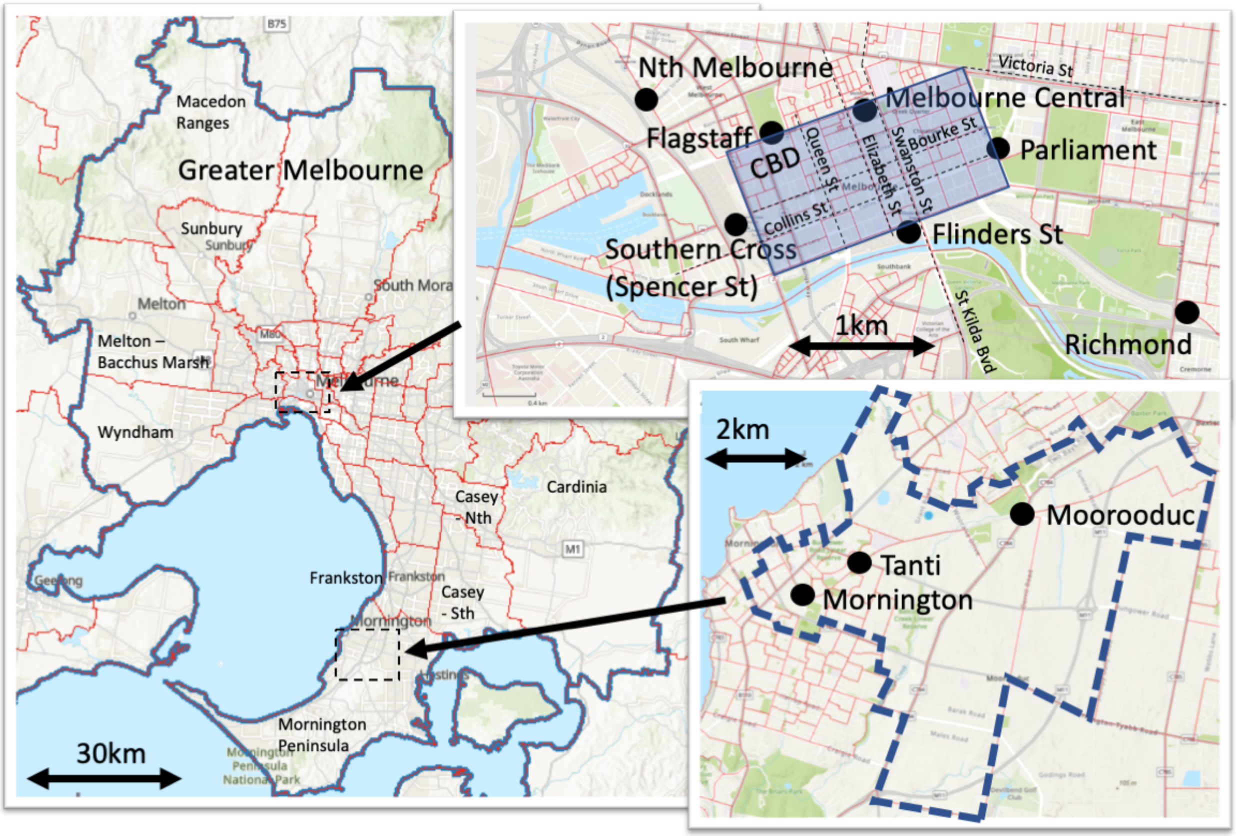


Figure 2: Areas of interest

3. Methodology

This study developed a package with tools for calculating the SI from GTFS data. The R programming language (R Core Team, 2023), was adopted for code development, and the package development setup and workflow described by Wickham and Bryan (2023) was adopted. Various existing packages were relied upon including: the sf package (Pebesma, 2023) for geospatial analysis; the tidyverse (Wickham et al., 2019); gtfstools (Herszenhut et al., 2022); and tidytransit (Poletti et al., 2023). Some code was adapted from examples, vignettes and other documentation in the tidytransit, gtfstools and other packages.

Two cases where used during the code development and testing such that results might be generated for real GTFS data: the Mornington Peninsula Tourist Railway GTFS feed and the Public Transport Victoria (PTV) GTFS feed, both in Victoria, Australia. Both were selected primarily for convenience, given that the authors are familiar with the typical service patterns and geography.

Figure @ref(Melbourne_map)) shows the areas of interest for which results are presented in this paper, including Greater Melbourne and its SA3 zones (main), SA1 zones in the central part of Melbourne (top-right) and the Mornington Peninsula Railway and SA1 zones within 800 metres (bottom-right) . Stations are shown in

Further cases were selected as leading, representative and contrasting examples for the results reported here.

3.1. Mornington Penninsula Tourist Railway

The Morning Peninsula Tourist Railway is in the outer south-east of Melbourne, running on Sundays and Wednesdays between Mornington and Moorooduc, with an intermediate stop at Tanti Park (see

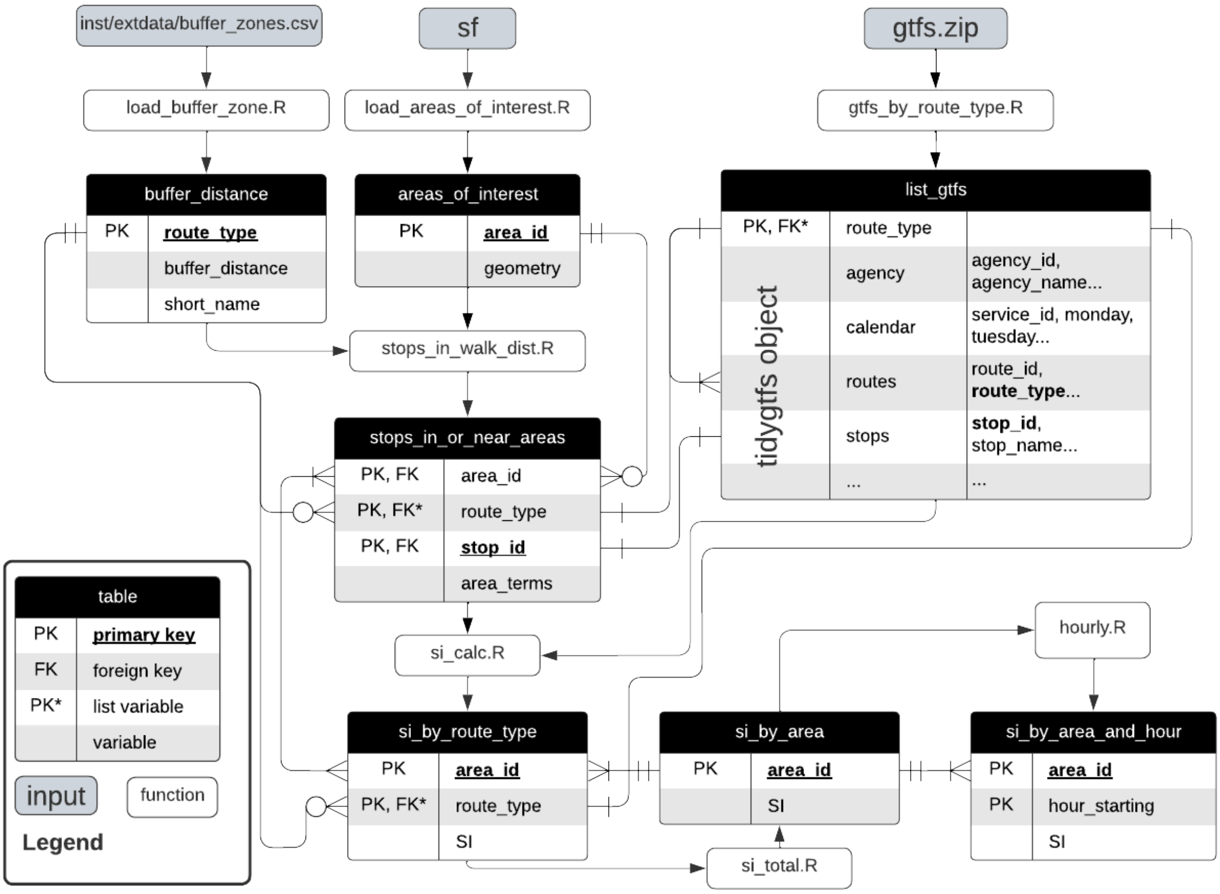


Figure 3: Entity Relationship Diagram (ERD) showing the data structure and functions related to the gtfssupplyindex package

https://transitfeeds.com/p/mornington-railway/806/latest/stops). A GTFS feed from 2018 was selected for the purposes of tests and demonstrating the code and output. Australian Bureau of Statistics (ABS) data was also used, sources via the strayr and absmapsdata packages (Mackey et al., 2023). The Mornington Peninsular Statistical Area 3 (SA3) zone and the Statistical Area 1 (SA1) zones contained within it were adopted as the areas of interest. These are shown in Figure @ref(fig:mornington_map_ABS), together with the locations of the three railway stations.

3.2. Public Transport Victoria (PTV)

Larger scale testing was performed using the Victorian GTFS feed, published by Public Transport Victoria (PTV), sourced via Transit Mobility Data, (2023) for historical feeds. Again, ABS data was used as the areas of interest, focusing on SA1 zones within Greater Melbourne. Data was analysed for the second Tuesday in August each year, so as to match the typical date of the Australian census.

4. Results

4.1. Code structure

Developed code is available and documented on github (Reynolds, 2024). The structure of the package, functions developed, and data tables are shown in Figure @ref(fig:SI_ERD), which shows
how the package takes input from three files: a gtfs feed (gtfs.zip); a sf object describing the geometry of the areas for which the SI is to be calculated; and a csv file defining the buffer zone distances (in metres)

for each route type[8]. The ultimate output is a si_by_area_and_hour table, which reports the SI score for each hour of the day across dates specified by the user.

4.2. Mornington Pennisula Tourist Railway

The various functions and their output and explained in the following, using the Mornington Peninsula GTFS for December 30th, 2018, and SA1 zone boundaries. Individual steps are:

(1) loading the gtfs.zip file: the gtfs_by_route_type function loads the gtfs data and splits it into a list (by route_type) of tidygtfs objects, using the filter_by_route_type function from the gtfstools package (Herszenhut et al., undated).
(2) loading geometry information about the areas of interest: geographical data about the areas of interest are loaded by the load_areas_of_interest.R function into an sf object, using the sf package (Pebesma, 2023). The resultant areas_of_interest table contains each area_id and its associated geometry. Data about buffer zones, specifically the walking distance threshold assigned to each route_type (mode) is then loaded, again through a function (load_buffer_zone.R).
(3) calculating which stops are within the catchment walking distance of which areas: using the stops_in_walk_dist function. Figure @ref(fig:calculate_stop_in_or_near_areas_verbose)) shows an intermediate step in this is function, in which the SA1 areas within the 800 metre catchment of the Mornington stations is identifed.

(4) Calculating SI scores for a given time period: using the si_calc.R function. This adapts code from an article included in the tidytransit package (Poletti, undated) to calculate the number of arrivals in a given time period, and combines this with the calculated area components. The si_total.R and hourly.R functions provided aggregation, giving the results shown in Table @ref(tab:SI_mornington_20181230_outpu and mapped in Figure @ref(fig:SI_mornington_20181230_output).

4.3. Central Melbourne

Figure @ref(Melbourne_CBD_map_230808) shows SI scores for Tuesday August 8, 2023, by hour between 5am and 11am (top) and by mode for the whole day (bottom). These generally expectations, with higher SI scores shown in the Central Business District (CBD), where there are the five stations that make up the City Loop: Flinders Street Station, Southern Cross Station, Flagstaff Station, Melbourne Central Station and Parliament Station; and where many tram and bus routes converge. The SI scores are highest between 7-9am, reflecting the typical service peaks. Results are also consistent with: the high number for bus services along the Victoria and Queen Street corridors; and tram services that mostly run along the Swanston, Elizabeth, Bourke and Collins Street corridors.

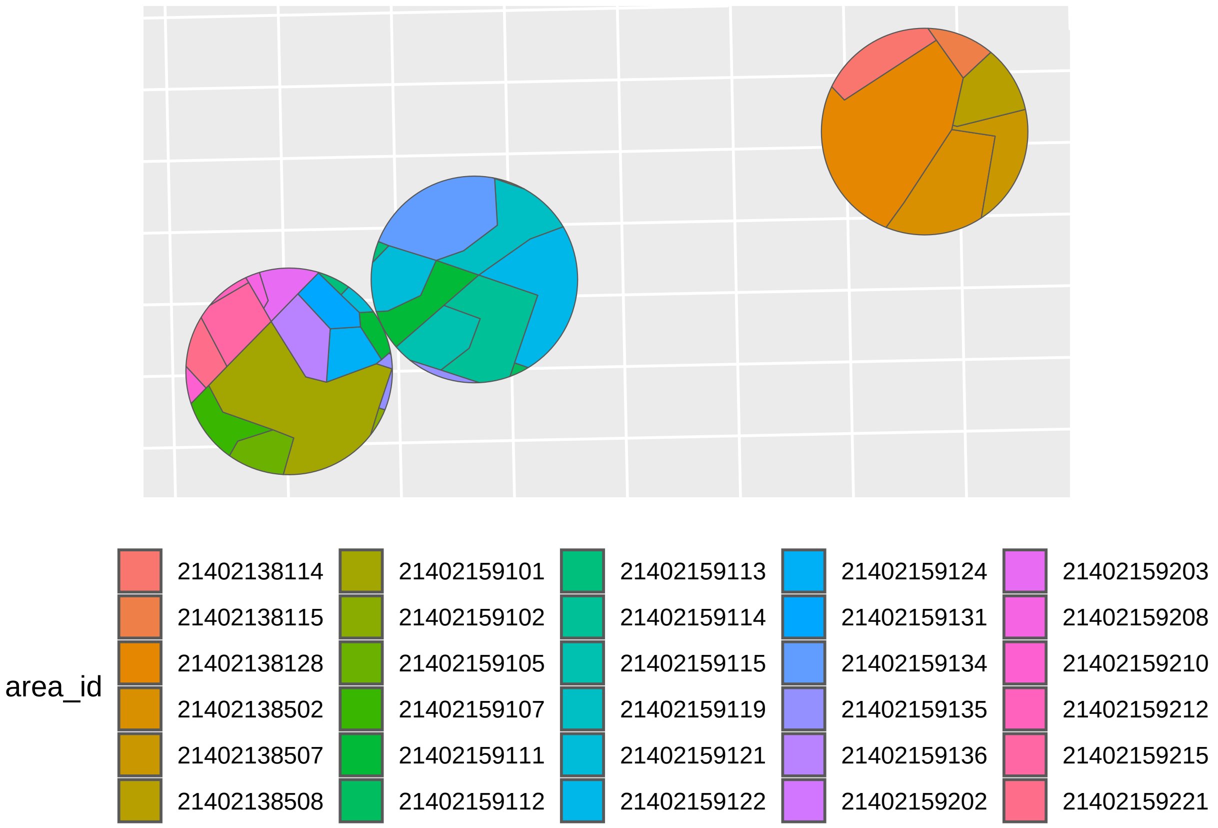


Figure 4: Step 3, stop catchments for the Mornington Pennninsula Tourist Railway, showing intersections with SA1 zones

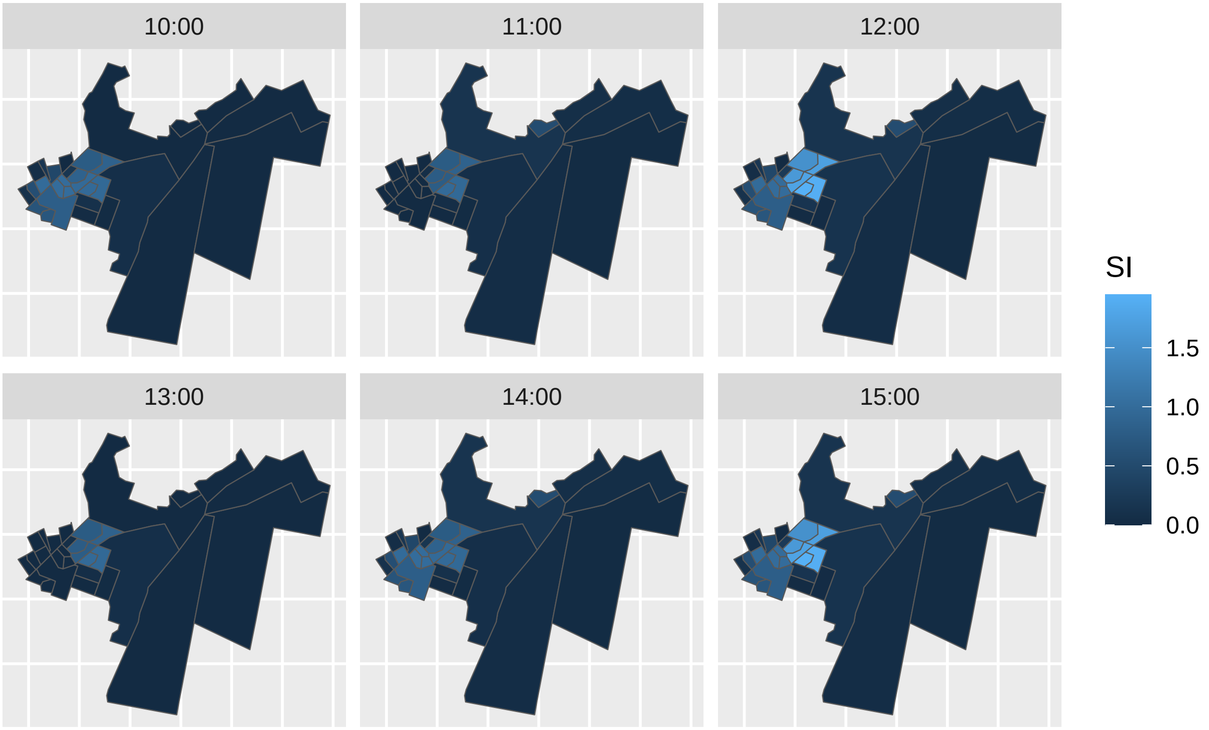


Figure 5: Mornington Penninsula Tourist Railway hourly SI values for December 30, 2018

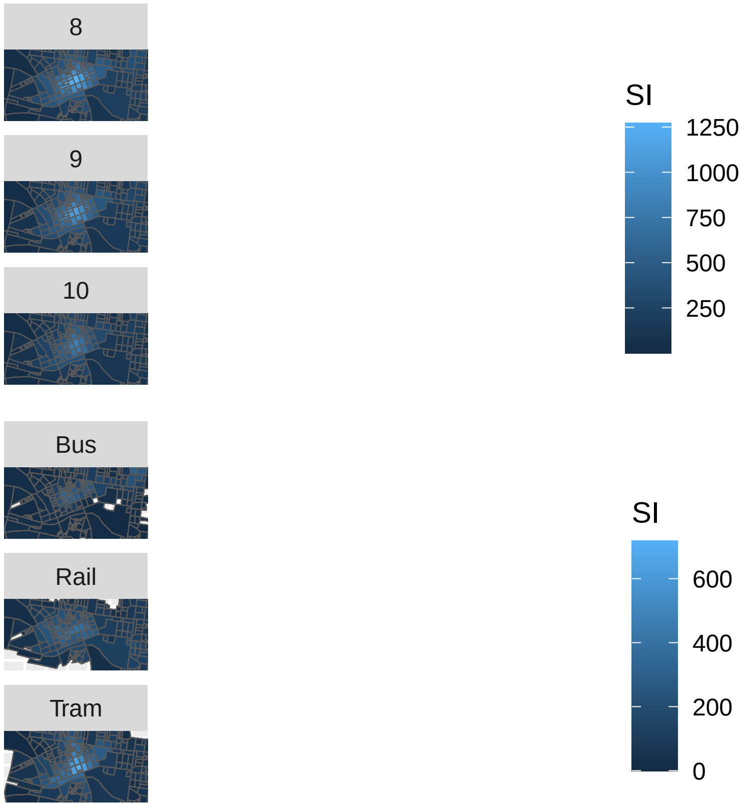


Figure 6: Victorian GTFS and central Melbourne SA1 zones, SI values for October 10, 2023, by hour between 5am and 11am (top) and by mode (bottom)

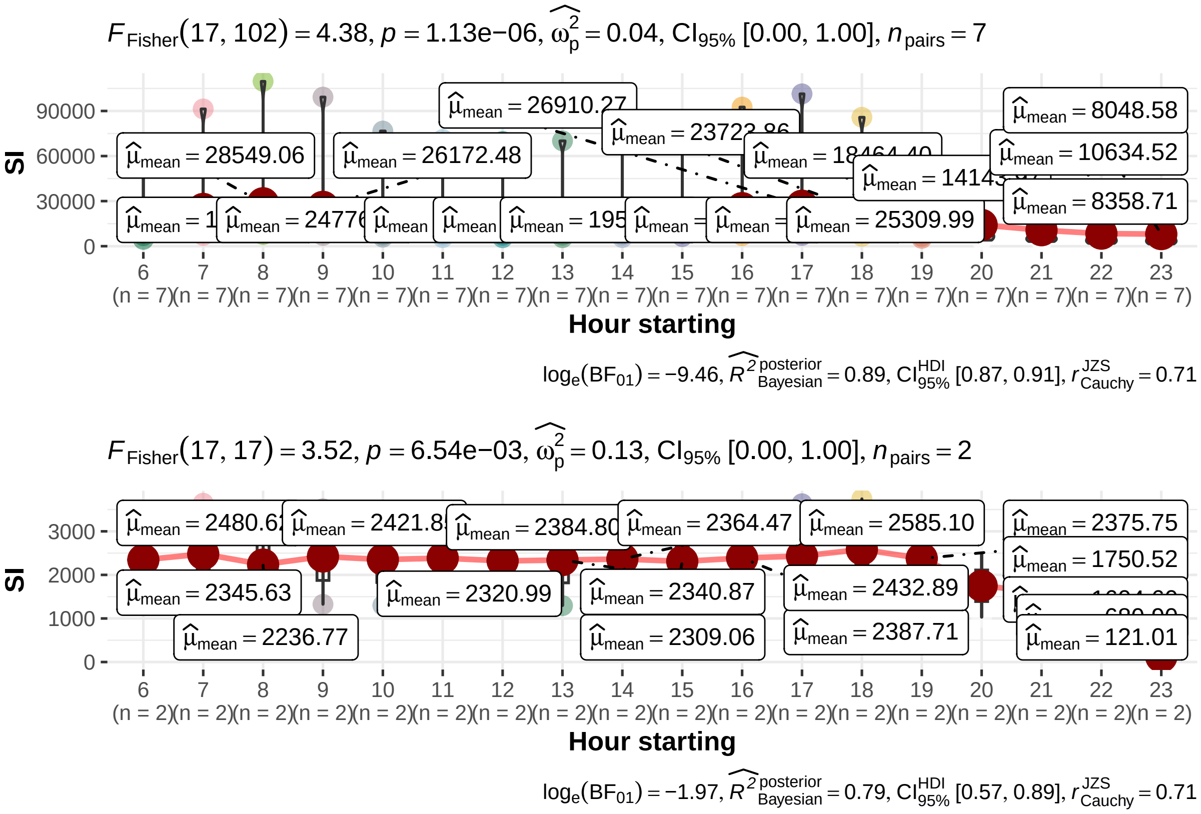


Figure 7: Victorian GTFS and SA3 zones, SI values for Tuesday August 8, 2023, by hour, 5am to midnight

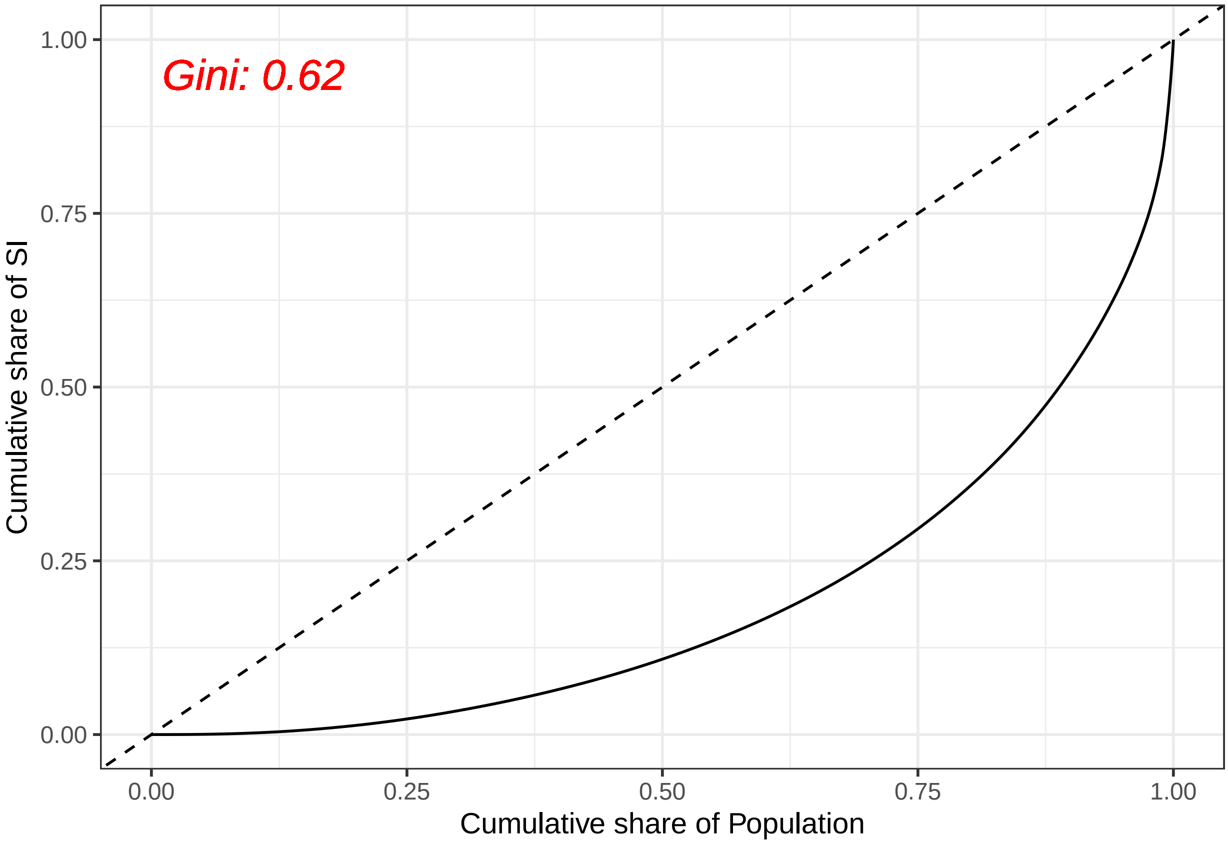


Figure 8: Victorian GTFS and SA1 zones within Greater Melbourne, SI values for Tuesday August 8, 2023, Lorenz plot

4.4. Greater Melbourne

4.4.1. By hour

4.4.2. Location, population and equality of service

Characteristic	**N = 40**
Population	
Median (IQR)	104,332 (77,574, 156,285)
Skew	1
Mean (SD)	118,389 (59,360)
10%, 90%	52,168, 193,124
5%, 95%	41,080, 218,426
1%, 99%	25,502, 264,363
Range	25,146, 279,213
SI	
Median (IQR)	94,230 (53,209, 136,011)
Skew	5
Mean (SD)	132,633 (208,439)
10%, 90%	17,527, 220,764
5%, 95%	12,483, 289,046
1%, 99%	5,282, 938,342
Range	839, 1,335,114

4.5. Trends

5. Extensions

5.1. Melbourne CBD Index

5.2. New York Index

5.3. London Index

6. Discussion and conclusions

References

Sultan Alamri, Kiki Adhinugraha, Nasser Allheeib, and David Taniar. Gis analysis of adequate accessibility to public transportation in metropolitan areas. *ISPRS international journal of geo-information*, 12(5):180, 2023. ISSN 2220-9964.

Robert A. Berenson. If you can't measure performance, can you improve it? *JAMA*, 315(7):645–646, 02 2016.

Felix Creutzig, Aneeque Javaid, Zakia Soomauroo, Steffen Lohrey, Nikola Milojevic-Dupont, Anjali Ramakrishnan, Mahendra Sethi, Lijing Liu, Leila Niamir, Christopher Bren d'Amour, Ulf Weddige, Dominic Lenzi, Martin Kowarsch, Luisa Arndt, Lulzim Baumann, Jody Betzien, Lesly Fonkwa, Bettina Huber, Ernesto Mendez, Alexandra Misiou, Cameron Pearce, Paula Radman, Paul Skaloud, and J. Marco Zausch. Fair street space allocation: ethical principles and empirical insights. *Transport Reviews*, 40(6):711–733, 2020. doi: 10.1080/01441647.2020.1762795. URL `https://doi.org/10.1080/01441647.2020.1762795`.

Graham Currie and Zed Senbergs. Identifying spatial gaps in public transport provision for socially disadvantaged Australians: the Melbourne needs-gap study. In *Australasian Transport Research Forum*. Australasian Transport Research Forum, 2007.

W Edwards Deming. The new economics for industry, government, education, 1993.

Gordon J Fielding. *Managing public transit strategically: a comprehensive approach to strengthening service and monitoring performance*. Jossey-Bass public administration series. Jossey-Bass Publishers, San Francisco, 1st ed. edition, 1987. ISBN 1555420680.

Florida Transit Information System. Urban integrated national transit database, 2018. URL `http://www.ftis.org/urban_intd.aspx`.

Luis A. Guzman, Daniel Oviedo, and Carlos Rivera. Assessing equity in transport accessibility to work and study: The bogotá region. *Journal of transport geography*, 58:236–246, 2017. ISSN 0966-6923.

Daniel Herszenhut, Rafael H. M. Pereira, Pedro R. Andrade, and Joao Bazzo. *gtfstools: General Transit Feed Specification (GTFS) Editing and Analysing Tools*, 2022. URL `https://ipeagit.github.io/gtfstools/`. R package version 1.2.0, https://github.com/ipeaGIT/gtfstools.

[8] This file is included in the package.

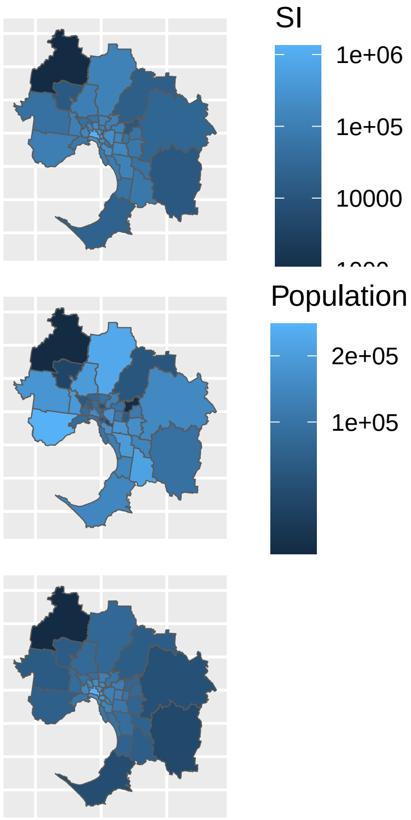


Figure 9: Victorian GTFS and SA3 zones within Greater Melbourne, SI values for Tuesday August 8, 2023 (top), 2021 census population (middle), and SI divided by population (bottom)

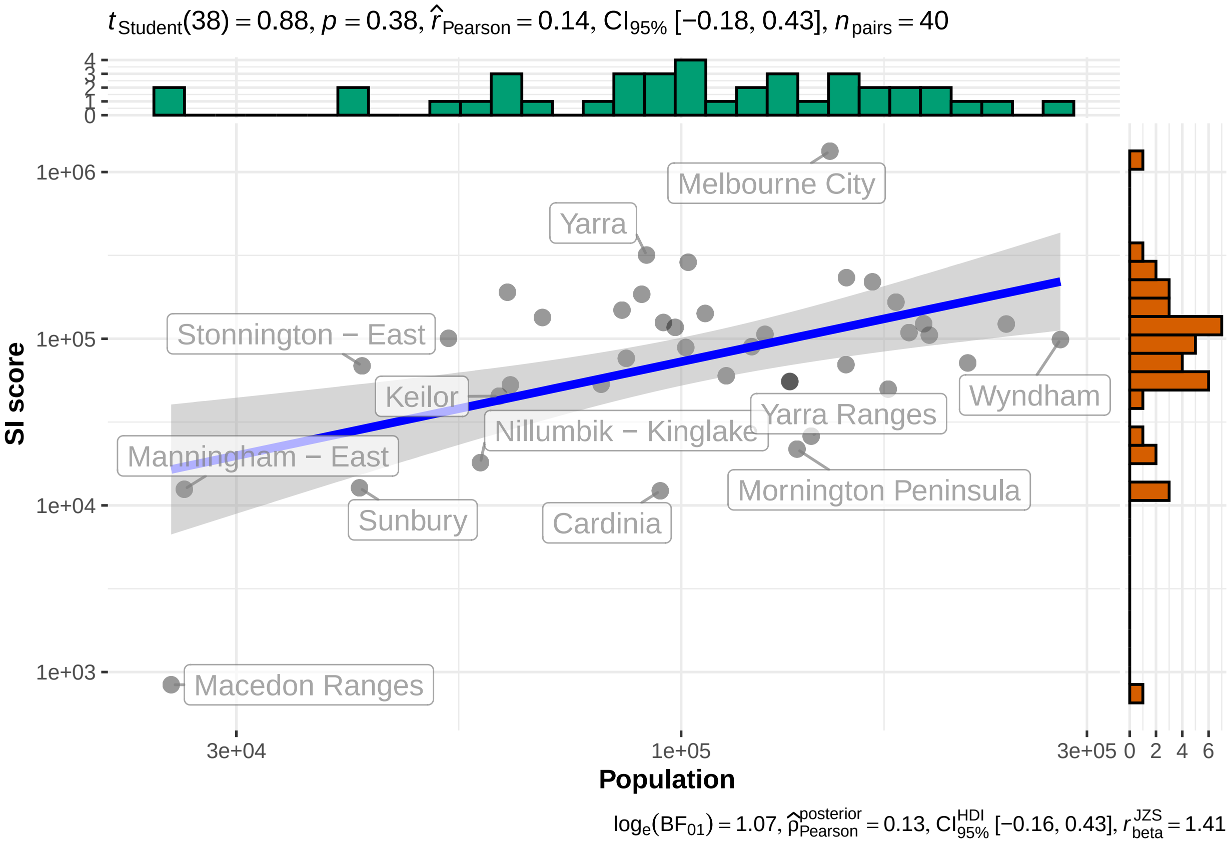


Figure 10: Victorian GTFS and SA3 zones within Greater Melbourne, SI values for Tuesday August 8, 2023 (top), 2021 census population (middle), and SI divided by population (bottom)

Danile Herszenhut, Rafael H. M. Pereira, Pedro R. Andrade, and Joao Bazzo. *gtfstools; filter GTFS object by route type (transport mode)*, undated. URL `https://ipeagit.github.io/gtfstools/reference/filter_by_route_type.html`. R package version 1.2.0.9000, last accessed June 30, 2023.

Imperial College London. Transport strategy centre (tsc); applied research, undated. URL `https://www.imperial.ac.uk/transport-engineering/transport-strategy-centre/applied-research/`.

International Association of Public Transport (UITP). Mobility in cities database 2015, 2015. URL `uitp.org/publications/mobility-in-cities-database/`.

Kittleson & Associates, Parsons Brinckerhoff, KFH Group, Texas A&M Transportation Institute, and ARUP. *Transit Capacity and Quality of Service Manual, Third Edition.* Transportation Research Board, Washington DC, third edition, tcrp report 165 edition, 2013. URL `http://www.trb.org/Main/Blurbs/169437.aspx`.

Todd Litman. Measuring transportation: traffic, mobility and accessibility. Technical Report 10, Institute of Transportation Engineers, Washington, D.C., 2003.

Todd Litman. When are bus lanes warranted? considering economic efficiency, social equity and strategic planning goals. Technical report, Victoria Transport Policy Institute, 2016. URL `http://www.vtpi.org/blw.pdf`.

Will Mackey, Matt Johnson, David Diviny, Matt Cowgill, Bryce Roney, William Lai, and Benjamin Wee. strayr, 2023. URL `https://runapp-aus.github.io/strayr/`.

MobilityData. *General Transit Feed Specification (GTFS)*, undated. URL `https://gtfs.org/`.

Edzer Pebesma. *sf: Simple Features for R*, 2023. URL `https://r-spatial.github.io/sf/`. R package version 1.0-14.

Flavio Poletti. *tidytransit: generate a departure timetable*, undated. URL `https://r-transit.github.io/tidytransit/articles/timetable.html`. R package version 1.5.0, last accessed June 22, 2023.

Flavio Poletti, Daniel Herszenhut, Mark Padgham, Tom Buckley, and Danton Noriega-Goodwin. *tidytransit: Read, Validate, Analyze, and Map GTFS Feeds*, 2023. URL `https://github.com/r-transit/tidytransit`. R package version 1.6.1.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL `https://www.R-project.org/`.

James Reynolds. gtfssupplyindex, 2024. URL `https://github.com/James-Reynolds/gtfssupplyindex`.

James Reynolds, Graham Currie, Geoff Rose, and Alistair Cumming. Moving beyond techno-rationalism: new models of transit priority implementation. In *Australasian Transport Research Forum 2017*, Auckland, New Zealand, 2017.

Paul Ryus, M Connor, S Corbett, A Rodenstein, L Wargelin, L Ferreira, Y Nakanishi, and K Blume. Tcrp report 88: a guidebook for developing a transit performance-measurement system. Technical report, 2003.

Transit Mobility Data,. Ptv gtfs - openmobilitydata, 2023. URL `https://transitfeeds.com/p/ptv/497`.

Walk Score. Transit score methodology. 2023. URL `https://www.walkscore.com/transit-score-methodology.shtml`.

Hadley Wickham and Jennifer Bryan. *R packages.* " O'Reilly Media, Inc.", 2023. URL `https://r-pkgs.org/`.

Hadley Wickham, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, Alex Hayes, Lionel Henry, Jim Hester, Max Kuhn, Thomas Lin Pedersen, Evan Miller, Stephan Milton Bache, Kirill Müller, Jeroen Ooms, David Robinson, Dana Paige Seidel, Vitalie Spinu, Kohske Takahashi, Davis Vaughan, Claus Wilke, Kara Woo, and Hiroaki Yutani. Welcome to the tidyverse. *Journal of Open Source Software*, 4(43):1686, 2019. doi: 10.21105/joss.01686.

James Wong. Leveraging the general transit feed specification for efficient transit analysis. *Transportation Research Record*, 1 (2338):11–19, 2013. doi: 10.3141/2338-02.