Leveraging GTFS data to assess transit supply

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

[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

While “if you can’t measure it, you can’t manage it” is often miss-attributed to Deming (1993), who was trying to make the opposite point (**?**), service level indicators are an important part of researching, managing and seeking to improve transit operations (Fielding, 1987; Ryus et al., 2003). A wide range of 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).

Practitioners, researchers and advocates using such metrics may face two inter-related challenges: (1) calculating the metrics themselves for a specific location and service pattern; and (2) explaining the metrics, their meaning and importance to those who might not be specialists in transit, such as to politicians or the general public. For example, the TCQSM metrics appear difficult to calculate in practice without access to specialist software and data. But, they appear relatively easy to explain given they use an A to F scoring system and there is an entire guidebook about them (although this may be offset by large number of indicators). In contrast, Transit Scores can be obtained simply by typing an address into a website, which will report a score out of 100 reflecting the quantity of transit available. However, Transit Scores cannot be calculated independently as the methodolgy / algorithm is not publicly available.

Previous research by Currie and Senbergs (2007) developed a transit Supply Index (SI) metric that appears to be both relatively easy to calculate and relatively simple to explain to non-transport professionals. It is obtained by calculating the number of transit arrivals at each stop within an area of interest, with an adjustment made to account for the typical walking distance catchment. Higher SI scores indicate areas with higher frequency 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 timetable data was not publicly available in a standardized and machine-readable format. The scores reported in Currie and Senbergs (2007) were calculated directly from a database of services provided by the transit authority in Melbourne, Australia. Since then, however, the General Transit Feed Specification (GTFS) has been developed as a way to publish timetable data in a standardized format.

[*] 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

More than 10,000 agencies are now providing GTFS feeds[4] (MobilityData, undated), and many visulization, procession and analysis tools are now available. A gap, however, is that there is not yet a tool to calculate SI scores directly from a GTFS dataset. 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 seciton 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 conclusion.

2. Background

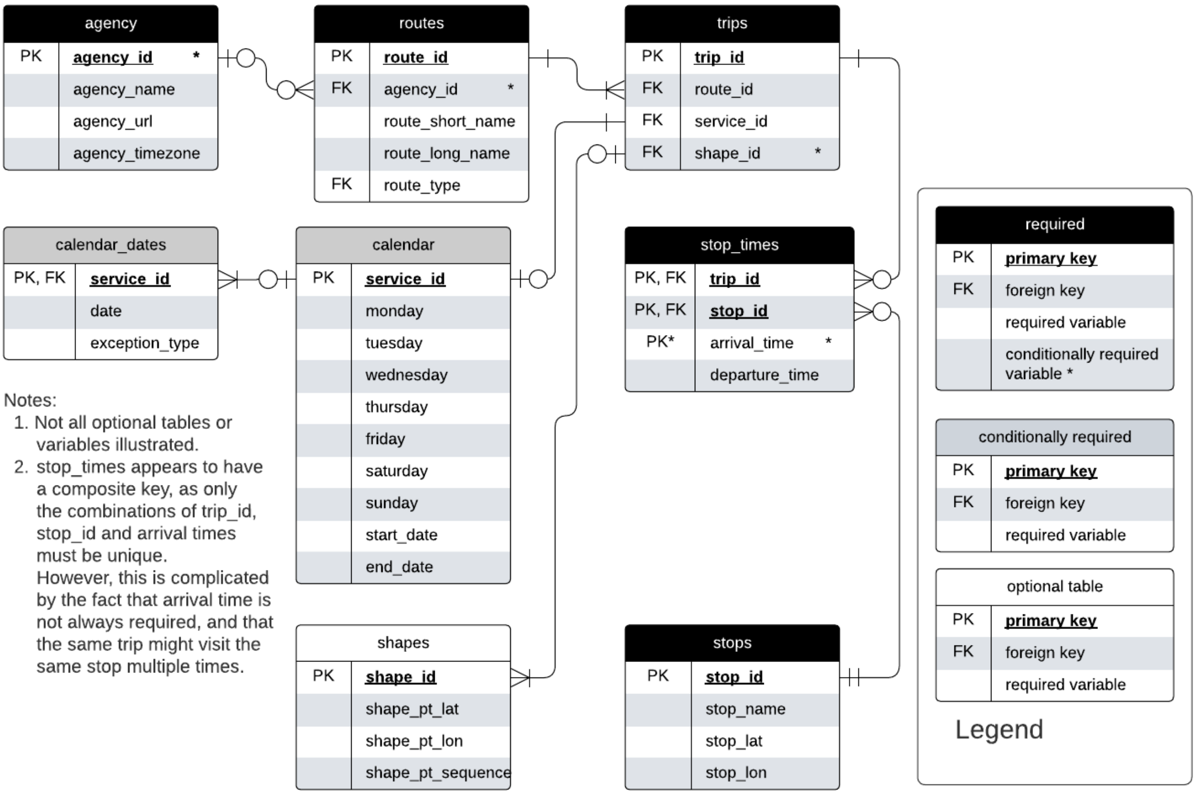
2.1. Transit metrics

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 programme undertaken yearly by the Transport Strategy Centre, which includes 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, service outputs and service consumption.
These can help describe cost efficiency, cost effectiveness or service effectiveness. At a larger scale, Litman (2003) and Litman (2016) discuss some of the traffic, mobility, accessibility, social equity, strategic planning and other rational decision-making frames that might underlie such transit metrics, while Reynolds et al. (2017) extends into models of how institutionalism, incrementalism and other public policy analysis concepts might apply to decision-making processes. Further examples include: (1) Guzman et al. (2017), who develop a measure of accessibility in the context of policy development and social equity for Latin American Bus Rapid Transit (BRT) networks; and (2) the street space allocation metrics based around 10 ethical principles introduced by Creutzig et al. (2020).

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 scores independently. Nor does it appear to be possible for Transit Scores to be generated for proposed changes to networks. Transit Score, therefore, fails the first of the aforementioned challenges, as practitioners, researchers and advocates can only use those scores provided online. The metric is simple to explain: the closer to 100, the better. However, 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 it is simple to understand, yet may not withstand scrutiny.

[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.



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.

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

Another example is the TCQSM, which specifies Levels of Service (LOS) between A and F across a range of factors[5]. This scoring scheme appears relatively simple to explain[6], and the detail within Kittleson & Associates et al. (2013) provides a resource for anyone wanting to better understand what the scores mean. However, calculation of many of TCQSM metrics may need specialised software and datasets[7] 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 introduction of the General Transit Feed Specification (GTFS) and widespread release of schedule data in this format, however, has helped towards making transit metrics more broadly available and usable. 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 1 shows an Entity Relationship Diargram (ERD) of the GTFS data structure. Each box represents a database table in the GTFS, with table rows indicating the variables (columns) included in each[8].

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

[6] A is good and F is bad. Also this scoring system matches the A to F LOS scoring used in many traffic capacity analysis software and manuals.

[7] 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.

[8] 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[9]. 'Crow's feet' indicate the relationships between each table[10].

GTFS now provides a mechanism for including individual transit systems in many online products and analyseses, 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[11]. While the Wong (2013) open-source code is readily available[12] this is now 11 years old and does not appear to be currently maintained. Future research may involve reviewing this code and using it to analyse modern GTFS feeds. However, 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

A generalized form of the Transit Supply Index (SI) is shown in Equation 1[13].

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

In Equation 1:

(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 distance of each stop. Hence, more services, more stops and higher frequencies increase the score. However, the SI does not incorporate service span, speed or other elements of a transit service. While these may be important to passenger experience, they might add considerable 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.

3. Methodology

This study involved the development of a package with tools for calculating the SI from GTFS data. R (R Core Team, 2023), a widely used and readily available statistical programming language, was adopted for code development. The package development setup and workflow described by Wickham and Bryan (2023) was adopted in this study. 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.

[9] For example, stop_id also appears in the 'stop_times' table as a Primary Key and Foreign Key.

[10] See https://i.stack.imgur.com/fxaAq.png for guide to the symbols. But, for example, the stops table is required, with the stop_id field providing a unique (primary) key for every stop. Within the stop_times table (which is also required) the stop_id field is a foreign key. Each unique stop_id can appear many times in the stop_times table, but must appear only once in the stops table. In the stop_times table each combination of trip_id, stop_id and arrival time must be unique (But, see note 2!) meaning that these fields represent a composite key.

[11] Daily average headways, route length and stop numbers for 50 transit operators.

[12] https://github.com/jcwong86/GTFS_Explore_Tool

[13] Currie and Senbergs (2007) focus was the context of Melbourne's Census Collection Districts (CCD) and calculations based on a week of transit service. CCDs predate the introduction of Statistical Areas 1, 2, 3, and 4 (SA1, SA2, SA3, SA4), and other geographical divisions currently used by the Australian Bureau of Statistics (ABS), which may be more familiar to readers from down under.

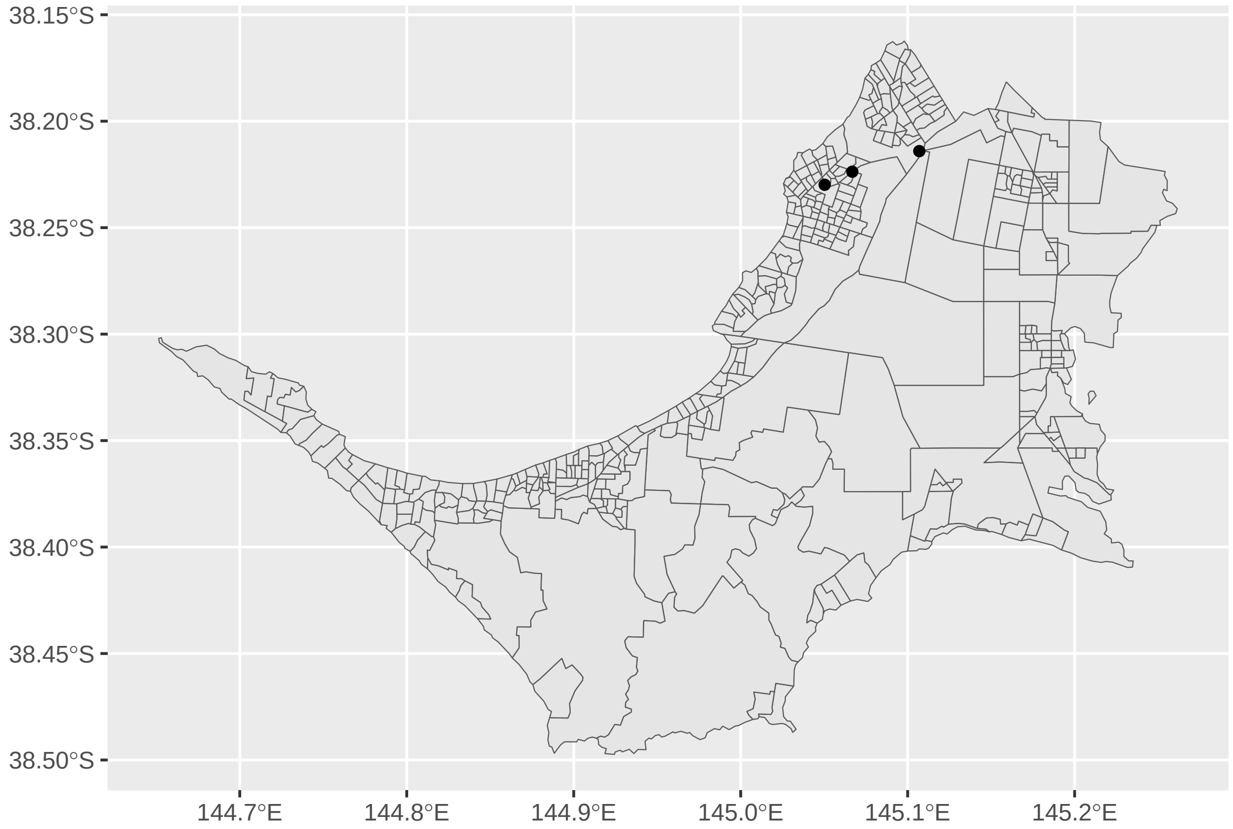


Figure 2: Mornington Penninsula SA1 zones and location of Mornington Tourist Rail stops.

Two cases where used as during the code development and testing so that results might be generated for real GTFS data. These were 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. Further cases were selected as leading, representative and contrasting examples for the results reported here.

3.1. Mornington Penninsula Tourist Railway

The Morning Penninsula Tourist Railway is located in the outer south-eastern suburbs of Greater Melbourne. It runs on Sundays and Wednesdays between Moorooduc and Mornington, with an intermediate stop at Tanti Park[14]. 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, primarily through 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 1, together with 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 for the areas_of_interest.

[14] https://transitfeeds.com/p/mornington-railway/806/latest/stops

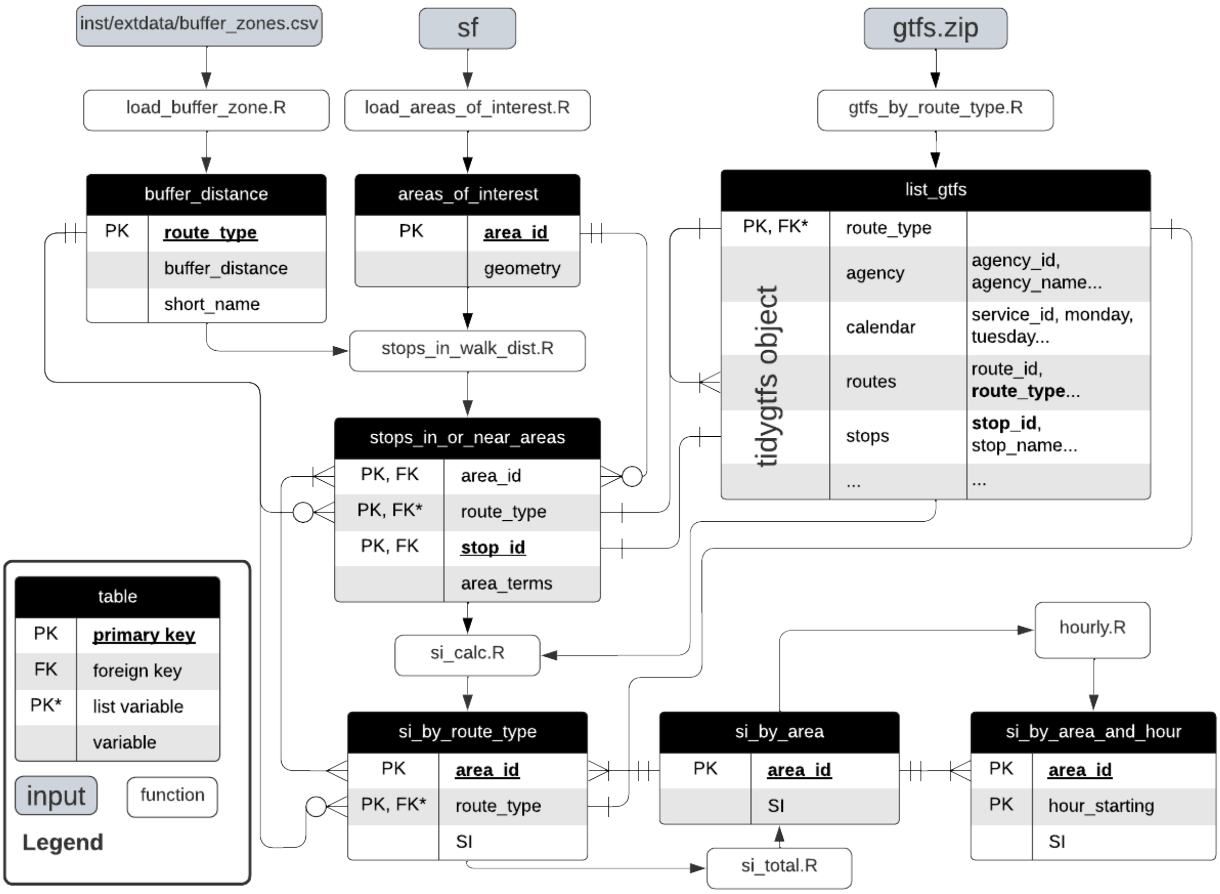


Figure 3: Entity Relationship Diagram (ERD) showing the data structure and functions related to the gtfssupplyindex package

3.3. Extensions??

Hourly "Manhattan- and London-ised Indexes"

Tidytransit includes a sample GTFS feed from New York's MTA (including the subway!), and so this was used for code tests were appropriate.

4. Results

4.1. Code structure and output

Developed code is available and documented on github (Reynolds, 2024). The structure of the package and the functions developed to generate each table are shown in Figure 2. This indicates 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[15].

Various data tables are output by functions included in the package. The ultimate output (Figure 2, bottom right) 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.

[15] A version of this file is included in the package.

Table 1: Mornington Penninsula Tourist Railway hourly SI values for December 30, 2018, for SA1 zones

area_id	10:00	11:00	12:00	13:00	14:00	15:00
214021381	0.0000672	0.0522962	0.0523635	0.0000672	0.0522962	0.0523635
214021385	0.0000000	0.0067970	0.0067970	0.0000000	0.0067970	0.0067970
214021591	0.2436873	0.1366432	0.3803305	0.1366432	0.2436873	0.3803305
214021592	0.0684965	0.0000000	0.0684965	0.0000000	0.0684965	0.0684965

4.2. Mornington Penninsula results

This section shows outputs of the various functions for December 30th, 2018, for the Mornington Peninsula Tourist Railway GTFS feed. The ultimate results, showing the SI scores for each Hourly results are shown in Table 2.

5. 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.

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.

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 1.0-15.

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.

Figure 4: Mornington Penninsula Tourist Railway hourly SI values for December 30, 2018, for SA1 zones

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.