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

The Transit Score website (Walk Score, 2023) will return a score out of 100 reflecting the quantity of transit available.
This score is easy to obtain and simple to explain to a non-technical audience. However, the score cannot be calculated independently, which 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.

In contrast, the Transit Capacity and Quality of Service Manual (TCQSM)(Kittleson & Associates et al., 2013) provides a 700 page guidebook full of indicators, all of which can be calculated independently if the necessary data is available. However, the TCQSM and many other transit service indicators developed in research and practice are focused towards the needs of practitioners and researchers. There is no simple 'overall' metric that might be easily obtained, used and/or understood by those who are not engineers, planners or other technically-minded people.

Previous research by Currie and Senbergs (2007) developed a transit Supply Index (SI), which has the advantage of providing a single score. The SI score is based on the number of transit arrivals at stops within an area of interest,
with an adjustment made to account for how much of the area of interest is within walking distance of each stop. 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 typically available in a standardized and machine-readable format.
Nowadays, the General Transit Feed Specification (GTFS) allows timetable publication in a standardized format, with more than 10,000 agencies providing feeds(MobilityData, undated). Many visualization, processing and analysis tools that accept GTFS data are now available, and it is this data that provides the input necessary for the Transit Score website to return a result.

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:

[*]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

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', beyond typical walking access distances to/from transit stops(Alamri et al., 2023).

The Fielding Triangle (Fielding, 1987) provides a framework for combining indicators of service inputs, outputs and consumption to describe cost efficiency, cost 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 transport indicators;
- 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 prioritization;
- 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; and
- 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 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. Transit Scores cannot be calculated independently or generated for proposed changes to networks.

2.2. GTFS

The General Transit Feed Specification (GTFS) is an open, text-based format developed originally to allow transit to be included in the Google Maps navigation platform (MobilityData, undated). Figure @ref(fig:GTFS_ERD) shows an Entity Relationship Diagram (ERD) of the GTFS data structure. This indicates how GTFS data is stored as a series of tables (agency, routes, trips etc.) with primary and foreign keys (agency_id, route_id, trip_id etc.) providing links.

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.

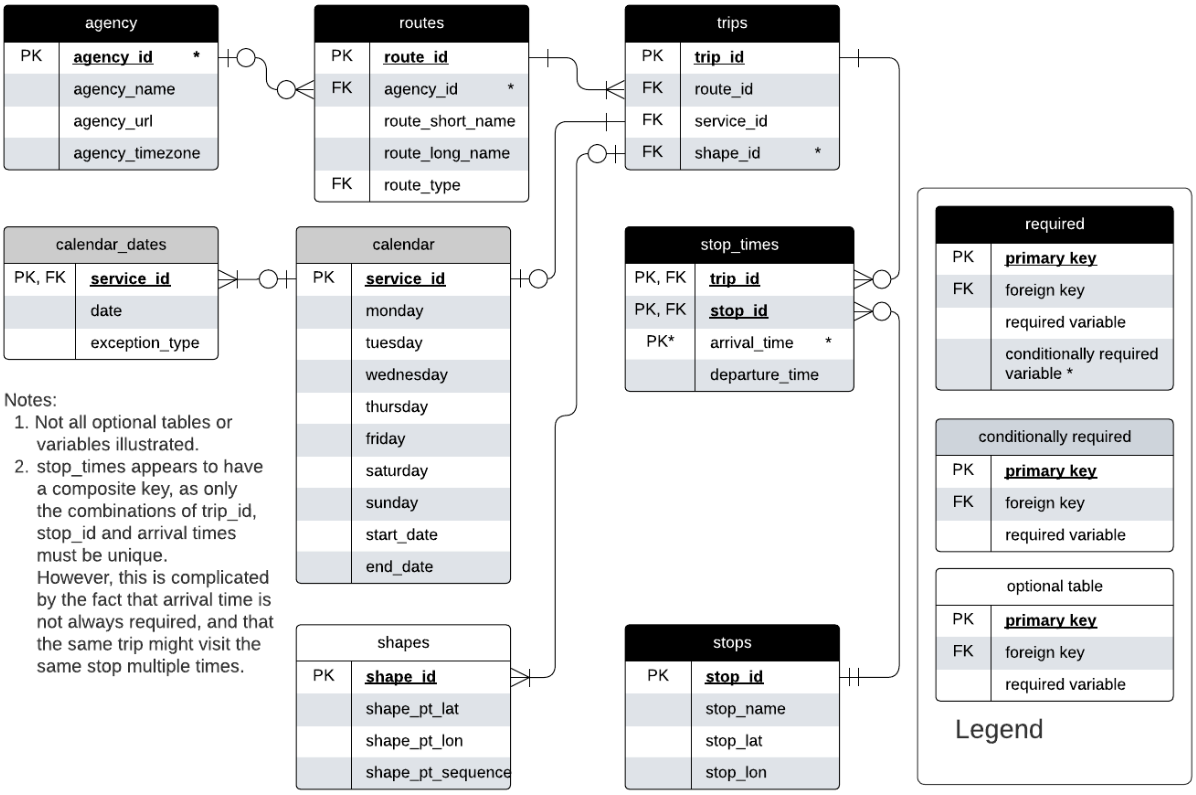

Figure 1: GTFS entity relationship diagram. Source: adapted by author from Alamri et al (2023) and the GTFS Schedule Reference (16/11/2023 revision).

2.3. The Transit Suppy Index

A generalized form of the Transit Supply Index (SI) from Currie and Senbergs (2007) is:

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

where: (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.

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.

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. Package development setup and workflow as 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 relevant to the code development and testing, and selected railway stations. Statistical Area (SA) zones from the Australian Bureau of Statistics (Australian Bureau of Statistics, undated) Areas of interest included Greater Melbourne and its SA3 zones (main), SA1 zones in the central part of Melbourne (top-right) and SA1 zones within 800 metres of the Mornington Penninsula railway (bottom-right).

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 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.

3.2. Public Transport Victoria (PTV)

Larger scale testing was performed using the Victorian GTFS feed, published by Public Transport Victoria (PTV) and sourced via Transit Mobility Data, (2023) for historical feeds.

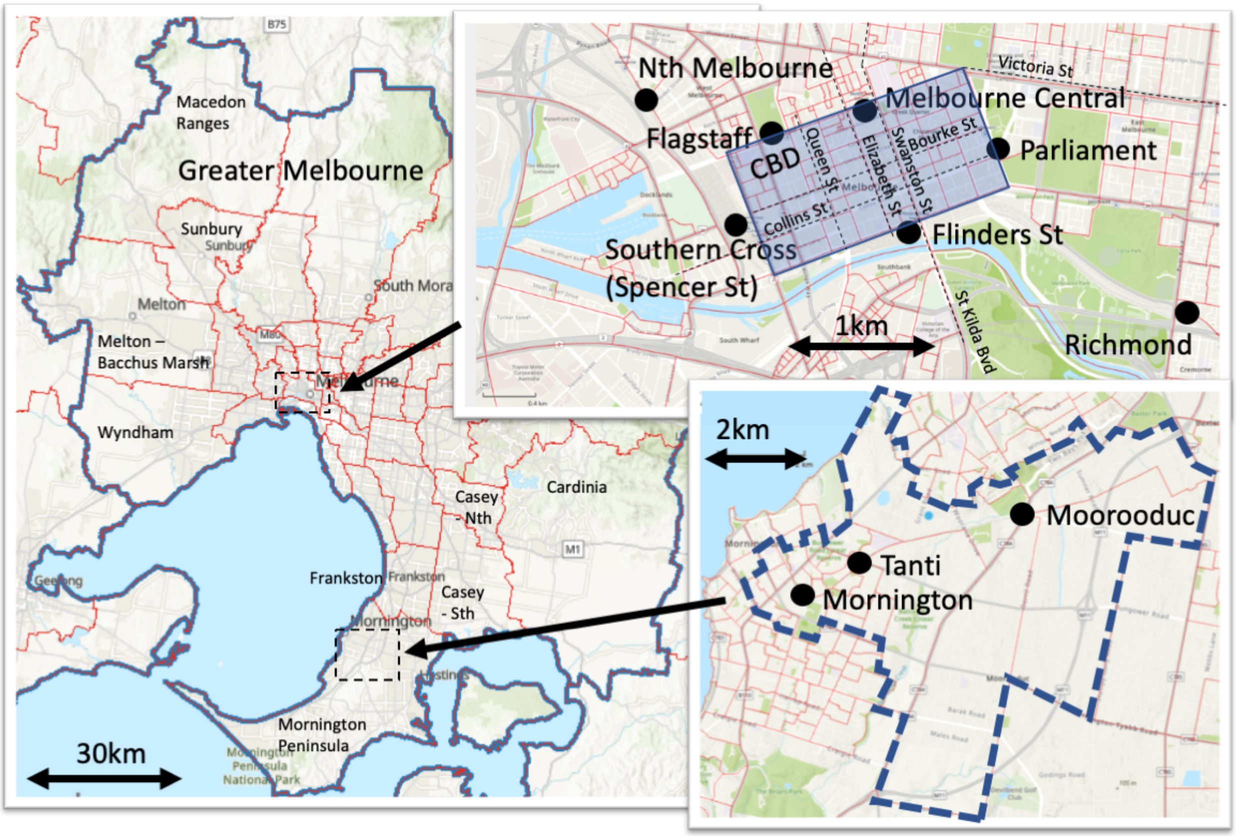

Figure 2: Areas of interest

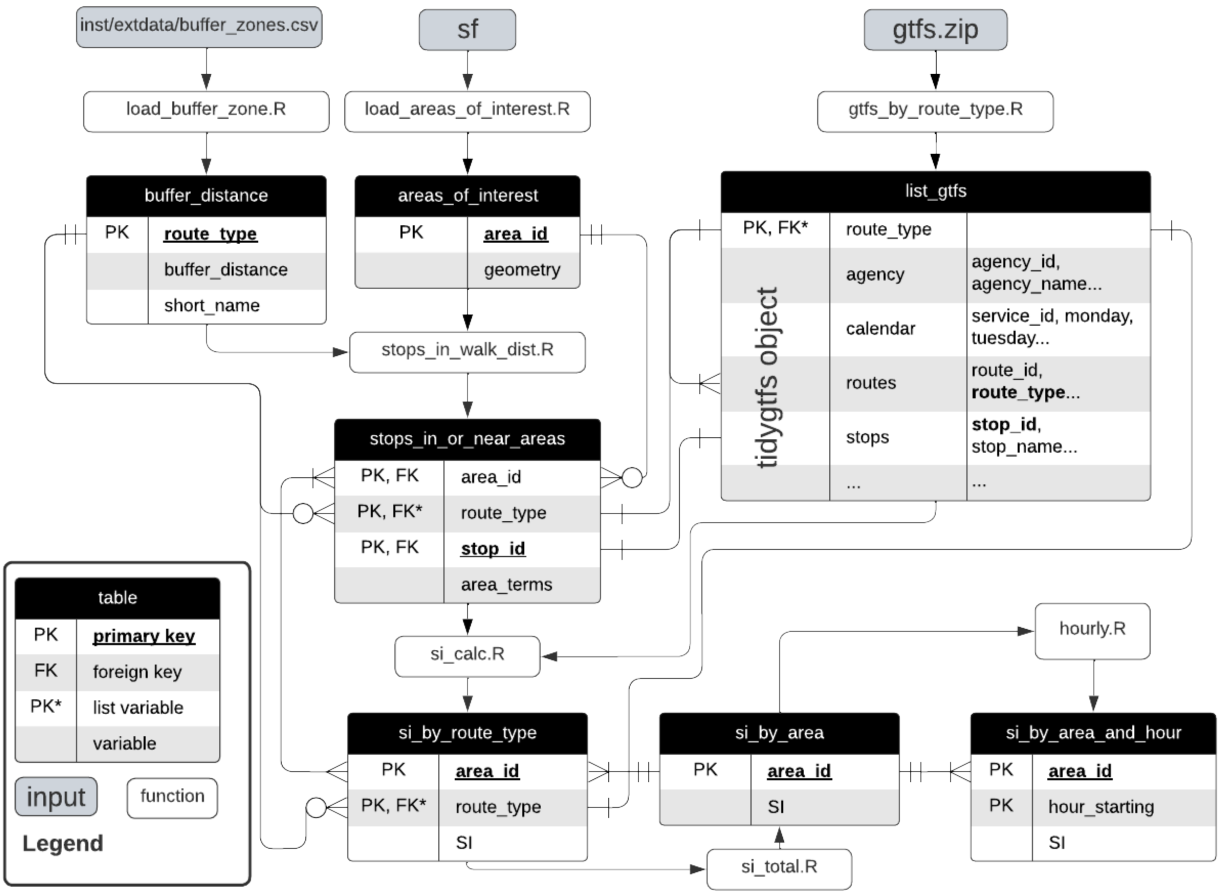

Figure 3: Entity Relationship Diagram (ERD) showing the data structure and functions related to the gtfssupplyindex package

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). This 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 (included in the package) defining the buffer zone distances for each route type. The ultimate output is a si_by_area_and_hour table (bottom-right), which reports the SI score for each hour of the day across dates specified by the user.

4.2. Mornington Pennisula Tourist Railway

Various functions and their output are explained in the following, using the Mornington Peninsula GTFS for December 30th, 2018, and SA1 zone boundaries as a worked example. 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

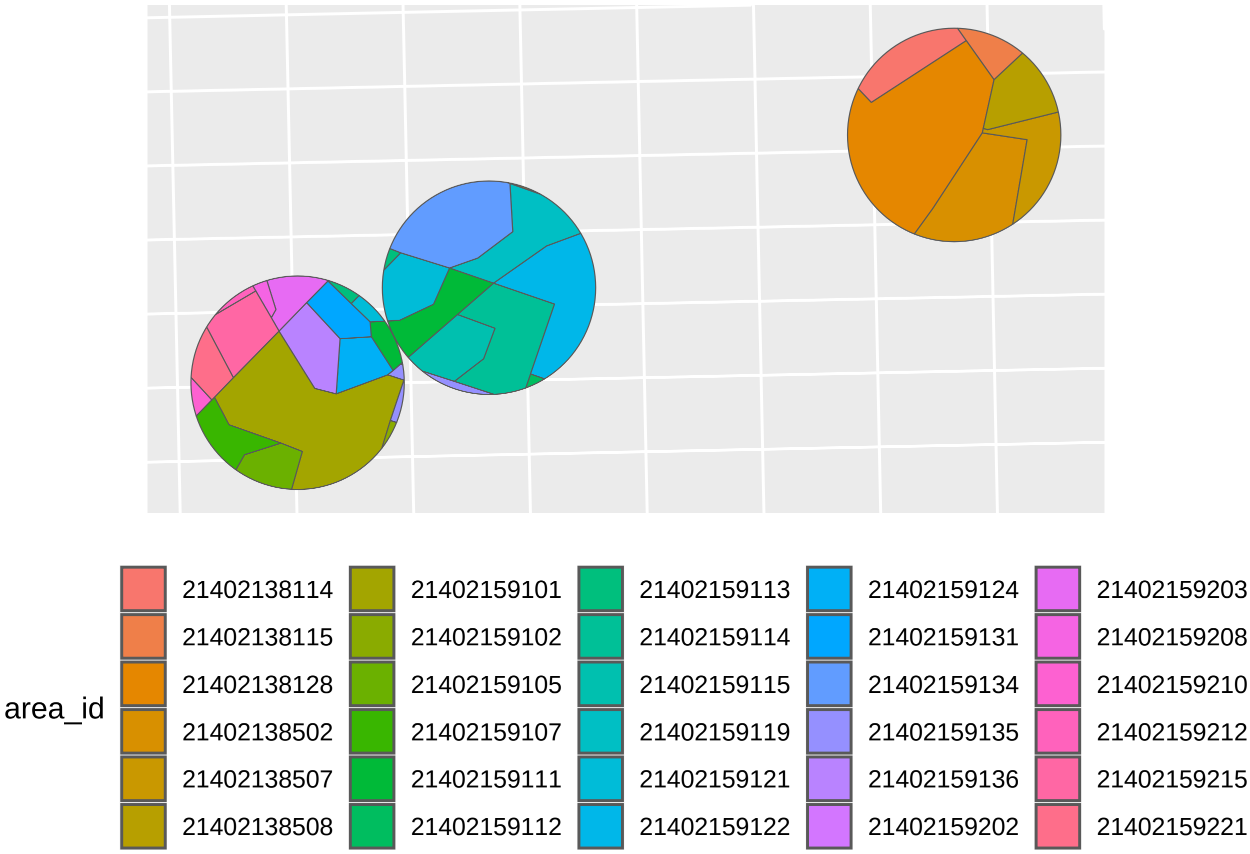

Figure 4: Step 3, stop catchments for the Mornington Penninsula Tourist Railway, showing intersections with SA1 zones

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 how this function identified SA1 areas within the 800 metre catchment of the three Mornington stations.

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

4.3. Central Melbourne

Figure @ref(fig:Melbourne_CBD_map_230808) SI scores for Tuesday August 8, 2023 by mode for SA1 zones in Central Melbourne. SI scores are generally higher in the Central Business District (CBD). Results are consistent with: the high number for bus services along the Victoria and Queen Street corridors; the rail stations surrounding the CBD; and tram services that mostly run along the Swanston, Elizabeth, Bourke and Collins Street corridors.

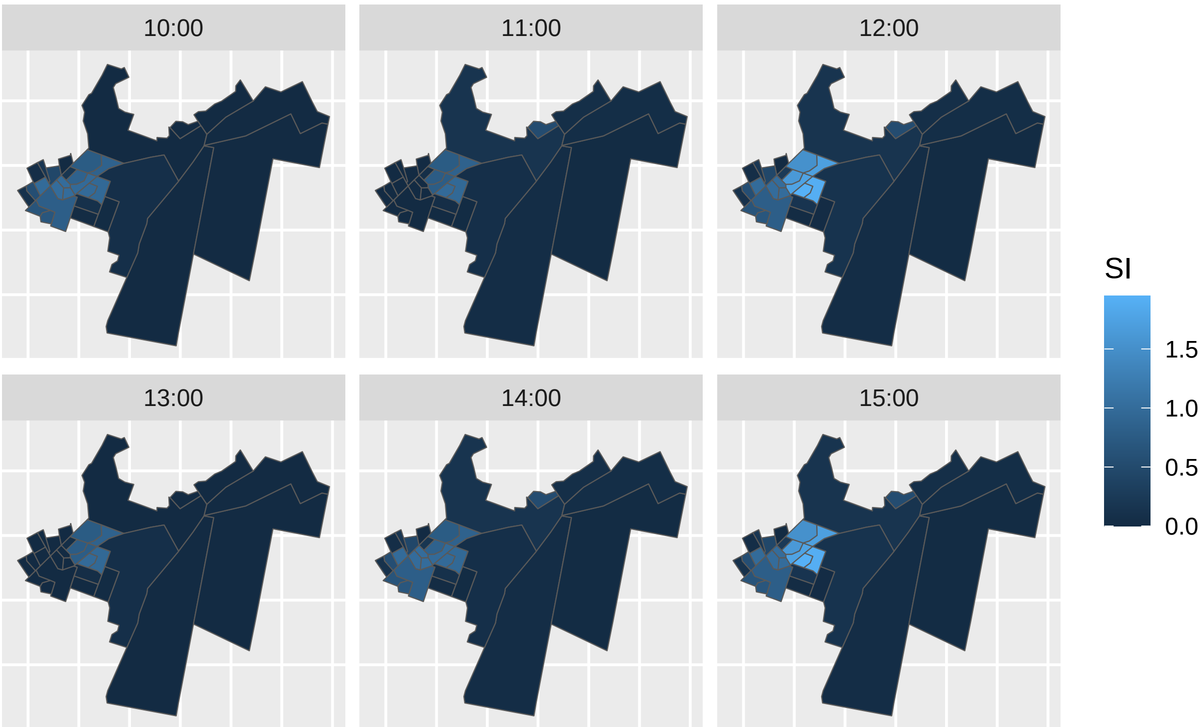

Figure 5: Mornington Penninsula Tourist Railway hourly SI values for December 30, 2018

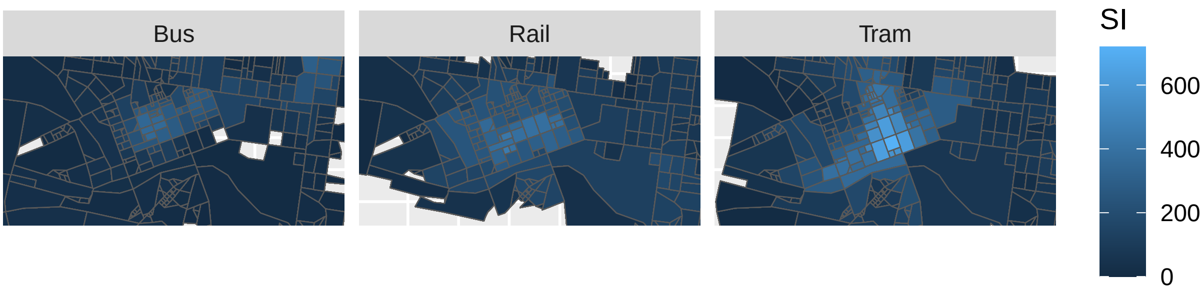

Figure 6: Victorian GTFS and central Melbourne SA1 zones, SI values for October 10, 2023, by mode

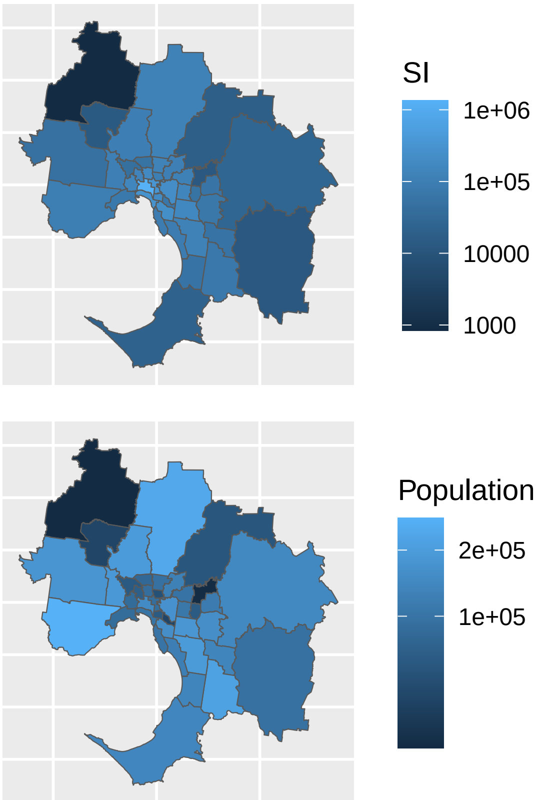

Figure 7: Victorian GTFS and SA3 zones within Greater Melbourne, SI values for Tuesday August 8, 2023 and 2021 census population

4.4. Greater Melbourne

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

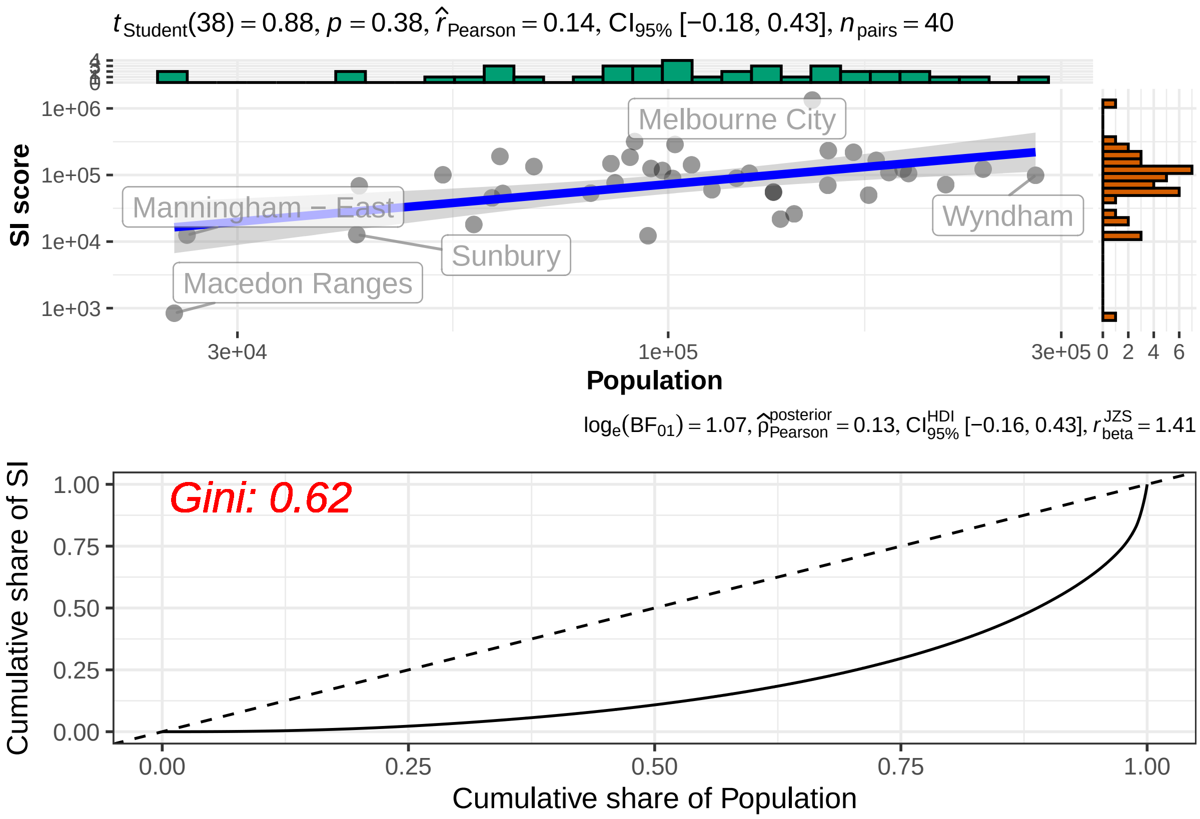

Figure 8: Victorian GTFS and SA3 zones within Greater Melbourne, SI values for Tuesday August 8, 2023 and 2021 census population

4.5. By year

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.

Australian Bureau of Statistics. Abs maps, undated. URL `https://maps.abs.gov.au/`.

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.

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.

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.