Journal Title
XX(X):2–9

DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

TTS2016R: A dataset to study population and employment patterns from the 2016 Transportation Tomorrow Survey (TTS) in the Greater Toronto and Hamilton Area, Canada

Anastasia Soukhov, Antonio Páez

Abstract

This paper describes and visualises the data contained within the {TTS2016R} data package created in R, the statistical computing and graphics language. In addition to a synthetic example, {TTS2016R} contains home-to-work commute information for the Greater Golden Horseshoe (GGH) area in Canada retrieved from the 2016 Transportation Tomorrow Survey (TTS). Included are all Traffic Analysis Zones (TAZ), the number of people who are employed full-time per TAZ, the number of jobs per TAZ, the count of origin destination (OD) pairs and trips by mode per origin TAZ, calculated car travel time from TAZ OD centroid pairs, and associated spatial boundaries to link TAZ to the Canadian Census. To illustrate how this information can be analysed to understand patterns in commuting, we estimate a distance-decay curve (i.e., impedance function) for the region. The value in {TTS2016R} is that it is a growing open data product built on R infrastructure that allows for the immediate access of home-to-work commuting data alongside complimentary objects from different sources. The package will continue expanding with additions by the authors and the community at-large by requests in the future. {TTS2016R} can be freely explored and downloaded in the associated Github repository where the documentation and code involved in data creation (including this manuscript), manipulation, and all open data products are detailed.

Prepared using sagej.cls

Keywords
Jobs; population; work; commute; travel time; impedance; Greater Toronto and Hamilton Area; Ontario, Canada; R

Introduction

This manuscript presents the open data product {TTS2016R}. Open data products are the result of turning source data (open or otherwise) into accessible information that adds value to the original inputs (see Arribas-Bel et al., 2021). The product presented in this paper is a R data package which currently consists of a fusion of objects from a variety of sources: home-to-work flows sourced from the 2016 Transportation Tomorrow Survey (TTS) (Data Management Group, 2018b), estimated travel times (calculated using {r5r} (Pereira et al., 2021)), and boundary files from the TTS (Data Management Group, 2018a) and from the Canadian Census (Statistics Canada).

What is a R data package? A R data package contains code, data, and documentation in a standardised collection format that can be installed by R users through a centralized software repository such as CRAN (the Comprehensive R Archive Network) and GitHub. {TTS2016R} is freely available on GitHub for all to install and freely use in the spirit of open and reproducible research. Currently and in more detail, {TTS2016R} includes full-time home-based work-to-job origin destinations (OD) counts and mode-specific trip numbers retrieved from the 2016 TTS, traffic analysis zone (TAZ) boundaries, and municipality, planning, and census metropolitan area boundaries for the Greater Golden Horse area (GGH) located in southern Ontario, Canada. In addition, the package includes TAZ centroid-to-centroid travel times by car, transit, cycling, and walking mode computed using package {r5r} (Pereira et al., 2021).

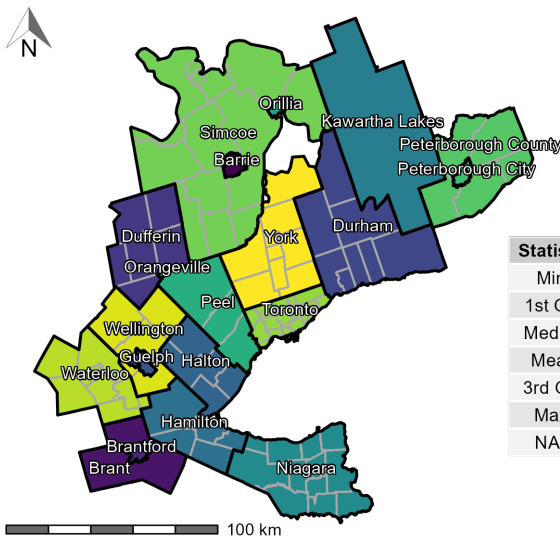
The aim of this paper is to walk readers through the data sets, illustrate a use case (i.e., the calculation of an impedance function that can be used to calculate accessibility to employment), and invite others to experiment in its uses and applications. Though data from the TTS is freely available to the public through the TTS Data Retrieval System, the raw data can be technically demanding, cumbersome to work with, and requires multiple software to process. By pre-processing the data, packaging it with complimentary data, and providing explicit documentation in a R environment, {TTS2016R} offers a slice of the TTS data that can be immediately used by R users to analysis patterns of commuting to work in the region. Anticipate this package to grow in the future: it currently provides an open infrastructure for additional TTS or complimentary data sets to be amended by the authors and the open-source community in the future by request.

Home-to-work commute data

Currently, {TTS2016R} includes counts of full-time employed population by place of residence (origin), counts of full-time usual place of work (destination), number of trips

to work by mode, and the calculated potential travel time of the trips in the GGH. The GGH (and hence the TTS survey area) is displayed in Figure 1.

This data is aggregated and available at the level of TAZ: TAZ are a spatial unit of analysis typically used to estimate the number of trips produced and attracted to each zone (Meyer and Miller, 2001). They are thus defined by transportation planners for a region based on intra-similarity and inter-dissimilarity between land-use and population demographics. Within the GGH boundaries, 3,764 TAZ are specified and each TAZ is uniquely identified using the GTA06 Zoning System: the survey boundary is discussed in the 2016 TTS methodology and defined by the TTS (Data Management Group, 2018b). The TAZ range between ≥ 0.019 km^2 in spatial area to a maximum of 879 km^2 (median: 1.3 km^2 and 3rd quantile: 2.8 km^2).



Statistic	Workers	Jobs	Links	TT
Min.	0	0	1	1
1st Qu.	25	64	14	13
Median	464	244	22	20
Mean	916	819	33	23
3rd Qu.	1378	700	38	30
Max.	8491	41821	1129	179
NA's	NA	NA	NA	3507

Figure 1. TTS 2016 study area within the GGH in Ontario, Canada along with associated descriptive statistic of workers and jobs per TAZ, OD links (count of workers potentially interacting with their place of employment) by origin TAZ, and calculated OD car travel time (TT) per origin TAZ. 3,507 trips were not assigned TT as they are longer than 180 mins. Spatial boundary files are retrieved from the TTS which define the survey area (Data Management Group, 2018a): the 20 regions in the GGH are represented by black lines and labelled, the dark gray lines are planning boundaries.

Full-time employed people and associated places of employment

In the GGH, there are 3,446,957 workers, 3,081,900 jobs, and 3,282,611 work-related trips (for the 2016 TTS survey day). The values are organized within the origin

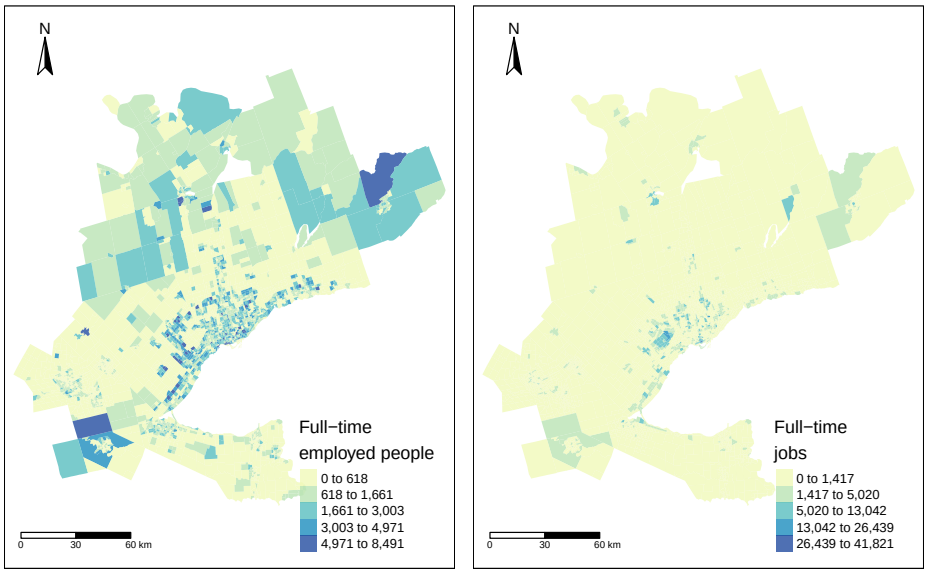


Figure 2. Number of workers (left) and jobs (right) in each TAZ retrieved from the 2016 TTS (Data Management Group, 2018b). Spatial boundary files are retrieved from the TAZ defined by the TTS (Data Management Group, 2018a).

destination (OD) table in the {TTS2016R} package and are derived from the cross-tabulation by person and by trip for the full-time employed population and associated places of employment. The TTS is a proportionally representative survey, hence the values included in {TTS2016R} are adjusted to reflect the GGH population.

It is important to note that the total number of full-time workers and jobs in the TTS 2016 region are not equal. Since the outer boundaries of the TTS are permeable, workers who reside within the boundaries but have workplaces that are outside of the boundaries are counted as workers within an origin TAZ, while jobs in TAZ that are filled by workers who reside outside the GGH boundaries are *unknown* since they were not surveyed. This mismatch results in the total number of workers being 1.12 times larger than the number of jobs (i.e., 3,446,957 workers to 3,081,900 jobs). As such, the OD table contained in {TTS2016R} offers a perspective on all workers in the GGH and their home-based trips to places of GGH employment.

The count of links and trips made by the full-time working population and associated full-time place of employment per unique OD pair are quite variable. TAZ contain between 0 to 8,491 workers (median: 464, 3rd quantile: 1,378), 0 to 41,821 jobs (median: 244, 3rd quantile: 700), and generate between 0 to 241 trips (median: 15, 3rd quantile: 42).

Figure 2 presents the number of employed people and associated jobs per TAZ. It can be observed that the spatial distribution of jobs and workers is unequal, which is

indicative of a jobs -housing imbalance that can impact accessibility in a region (Levine, 1998). It can also be seen that there is a higher number of TAZ with no workers than zones with no jobs (i.e., 791 TAZ with no workers : 396 TAZ with no jobs) and the mean of workers per TAZ is higher than the mean of jobs. The number TAZ with an extreme number of jobs at the highest and lowest percentiles is significantly higher than the number of workers.

Calculated travel time

Also included in {TTS2016R} are the estimated travel times between OD as summarized in descriptive statistics table in Figure 3; travel times are calculated using the package {r5r}. {r5r} interfaces with the java-based R5 routing engine developed separately by Conveyal (Conveyal, 2022-09-28T15:43:14Z). The inputs to {r5r} for this data package were the desired mode, a maximum travel time threshold of 180 minutes, the geo-coded origin destination pairs based on the centroids of the TAZ, and the static Open Street Map road network of Ontario (retrieved using Geofabrick (Geofabrik, 2022)). A travel time threshold of 180 minutes was selected since it captures almost all potential OD interactions.

Additionally, the car mode was included since it is a critically important commute mode in the GGH. 2,598,379 of the trips are made using a car mode out of the total 3,282,611 work-related trips according to the TTS 2016 data (i.e., 79% of trips are taken by car).

These travel times are useful addition to {TTS2016R} since they are not included in the TTS Data Retrieval System but they are vitally important to estimate the cost of travel and associated impedance functions, among other possible applications. If the readership is interested in additional information regarding the travel time computation, please see the calculation notebook in the documentation of {TTS2016R} and details about {r5r} at the r5r package website.

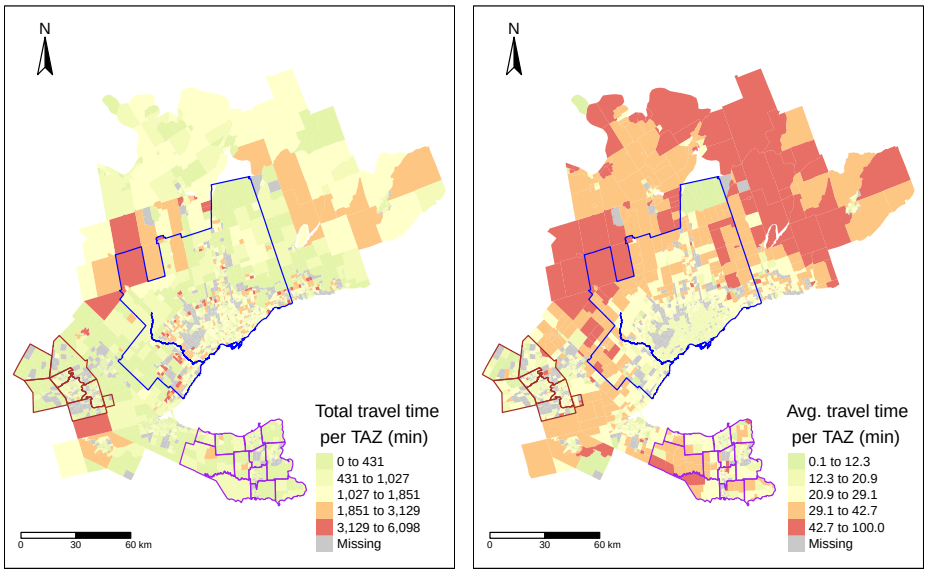


Figure 3. Calculated total worker travel time (left) and average worker travel time (right) for each TAZ in the 2016 TTS. Planning boundaries of Niagara and Waterloo (Data Management Group, 2018a), and the Toronto census metropolitan area (Statistics Canada, 2017) are drawn with purple, brown and blue borders, respectively.

As can be observed in Figure 3, the total travel time resembles the spatial trend distribution in the number of employed people in the previous plot (Figure 2) and the spatial distribution of the average travel time is distinct from other plots presented so far. For instance, we can see that in areas around the south-eastern border such as Niagara and Waterloo (purple and brown borders), the average travel times are moderately low. Additionally, travel times (by car) within the core of the Toronto census metropolitan area (CMA) (blue) is also moderately since traffic congestion is not reflected in the travel time estimations. Further from these areas, travel times are higher.

Calibrating an impedance function

Impedance functions are useful to understand mobility behaviour and are used to estimate gravity models of spatial interaction (Wilson, 1971; Haynes and Fotheringham, 1985) and applied in accessibility analysis (Hansen, 1959; Talen and Anselin, 1998; Páez et al., 2013; Barboza et al., 2021). An impedance function $f(\cdot)$ depends on the cost of travel c_{ij} between locations i and j (all which is supplied in the travel time and origin-destination table within {TTS2016R}).

A useful technique to calibrate an impedance function is to use the trip length distribution (TLD) as measured from origin-destination data (Horbachov and Svichynskyi, 2018; Batista et al., 2019). The TLD is the representation of the likelihood

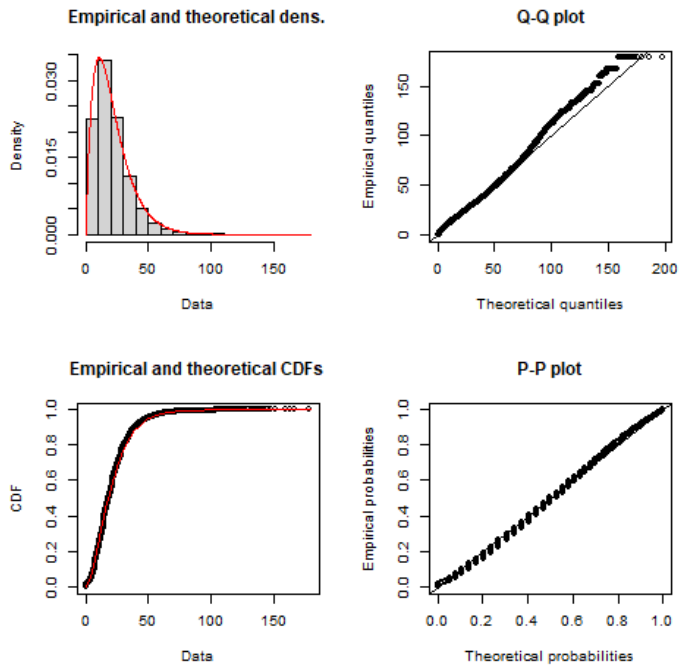


Figure 4. Empirical TTS 2016 home-based car TLD (black) and calibrated gamma distribution impedance function (red) with associated Q-Q and P-P plots

that a proportion of trips are taken at a specific travel cost. In our data set, where we assume cost is travel time, the impedance function maps low travel times to higher proportions of trips, and high travel times are mapped to low proportion of trips.

Using the data contained in {TTS2016R}, we fit the empirical TLD to a density distribution using maximum likelihood techniques and the Nelder-Mead method for direct optimization available within the R package {fitdistrplus} (Delignette-Muller and Dutang, 2015). Based on goodness-of-fit criteria and diagnostics seen in Figure 4, the gamma distribution is selected. The 'shape' parameter is α = 2.019, the estimated 'rate' is β = 0.094 , and $\Gamma(\alpha)$ is defined in Equation (1).

$$f(x,\alpha,\beta) = \frac{x^{\alpha-1}e^{-\frac{x}{\beta}}}{\beta^{\alpha}\Gamma(\alpha)} \quad \text{for } 0 \leq x \leq \infty$$
$$\Gamma(\alpha) = \int_0^{\infty} x^{\alpha-1}e^{-x}\,dx \tag{1}$$

\end{equation}

Concluding remarks

{TTS2016R}, the open data package introduced in this paper fuses multiple sources of data. It includes an OD cross-tabulation by person and by trip mode table for home-to-work commute data from the 2016 TTS alongside complimentary boundaries and estimated travel times. The value of this data package is in its transparency, easy of access, and its open infrastructure for the addition of complimentary data sets in the future. Using `R` users can immediately and easily explore GGH commute flow trends as well as suggestion further amendments to the package by request. One possible use of this data, as showcased in this paper, is the calibration of impedance functions which in turn can be used for accessibility analysis.

In the spirit of novel and original research, we hope readers value the efforts made to detail the data in order to improve transparency in our work and encourage others to replicate and, hopefully, inspire research of their own. We see this product as providing open infrastructure for additional TTS or complimentary data sets to be amended by the authors or wider open-source community in the future.

References

Arribas-Bel D, Green M, Rowe F and Singleton A (2021) Open data products-a framework for creating valuable analysis ready data. *Journal of Geographical Systems* 23(4): 497–514. DOI:10.1007/s10109-021-00363-5. URL https://dx.doi.org/10.1007/s10109-021-00363-5.

Barboza MHC, Carneiro MS, Falavigna C, Luz G and Orrico R (2021) Balancing time: Using a new accessibility measure in Rio de Janeiro. *Journal of Transport Geography* 90: 102924. DOI:10.1016/j.jtrangeo.2020.102924. URL https://www.sciencedirect.com/science/article/pii/S0966692320310012.

Batista S, Leclercq L and Geroliminis N (2019) Estimation of regional trip length distributions for the calibration of the aggregated network traffic models. *Transportation Research Part B: Methodological* 122: 192–217. DOI:10.1016/j.trb.2019.02.009. URL https://linkinghub.elsevier.com/retrieve/pii/S0191261518311603.

Conveyal (2022-09-28T15:43:14Z) Conveyal R5 Routing Engine. URL https://github.com/conveyal/r5.

Data Management Group (2018a) Survey Boundary Files. URL http://dmg.utoronto.ca/survey-boundary-files.

Data Management Group (2018b) TTS - Transportation Tomorrow Survey 2016. URL http://dmg.utoronto.ca/transportation-tomorrow-survey/tts-introduction.

Delignette-Muller ML and Dutang C (2015) fitdistrplus: An R package for fitting distributions. *Journal of Statistical Software* 64(4): 1–34. URL https://www.jstatsoft.org/article/view/v064i04.

Geofabrik (2022) Ontario OpenStreetMap - Geofabrik Download Server. URL https://download.geofabrik.de/north-america/canada/ontario.html.

Hansen WG (1959) How Accessibility Shapes Land Use. *Journal of the American Institute of Planners* 25(2): 73–76. DOI:10.1080/01944365908978307. URL http://www.tandfonline.com/doi/abs/10.1080/01944365908978307.

Haynes KE and Fotheringham AS (1985) *Gravity and Spatial Interaction Models*. Reprint. WVU Research Repository. URL https://researchrepository.wvu.edu/cgi/viewcontent.cgi?article=1010&context=rri-web-book.

Horbachov P and Svichynskyi S (2018) Theoretical substantiation of trip length distribution for home-based work trips in urban transit systems 11(1): 593–632. URL https://www.jstor.org/stable/26622420. Publisher: Journal of Transport and Land Use.

Levine J (1998) Rethinking accessibility and jobs-housing balance. *Journal of the American Planning Association* 64(2): 133–149. URL ISI:000073499600007. JSPR.

Meyer MD and Miller EJ (2001) *Urban transportation planning: a decision-oriented approach*. McGraw-Hill series in transportation, 2nd ed edition. Boston: McGraw-Hill. ISBN 978-0-07-242332-7.

Pereira RHM, Saraiva M, Herszenhut D, Braga CKV and Conway MW (2021) r5r: Rapid realistic routing on multimodal transport networks with r^5 in r. *Findings* DOI:10.32866/001c.21262.

Páez A, Farber S, Mercado R, Roorda M and Morency C (2013) Jobs and the Single Parent: An Analysis of Accessibility to Employment in Toronto. *Urban Geography* 34(6): 815–842. DOI:10.1080/02723638.2013.778600. URL http://www.tandfonline.com/doi/abs/10.1080/02723638.2013.778600.

Statistics Canada (????) Boundary Files, 2016 Census. URL https://www12.statcan.gc.ca/census-recensement/2011/geo/bound-limit/bound-limit-2016-eng.cfm.

Talen E and Anselin L (1998) Assessing Spatial Equity: An Evaluation of Measures of Accessibility to Public Playgrounds. *Environment and Planning A: Economy and Space* 30(4): 595–613. DOI:10.1068/a300595. URL http://journals.sagepub.com/doi/10.1068/a300595.

Wilson A (1971) A family of spatial interaction models, and associated developments. *Environment and Planning A: Economy and Space* 3(1): 1–32. DOI:10.1068/a030001. URL http://dx.doi.org/10.1068/a030001.