Reviewer Response Letter

The authors thank the reviewers for their time and thoughtful commentary in response to our paper. We have extensively revised the paper to clarify the comments in the paper. In this respect, we present the comments in blue text, our responses and comments in black text, and the relevant revised text is presented in red.

# Response to Referee #1:

## Firstly, I should say that I've never used R, so I can't comment on how unique or useful these techniques are in terms of the libraries being used. What I would like to have seen, though, is how computationally efficient this method is and how it would scale to larger datasets (e.g. memory requirements or time to compute). There is no mention of how long the cost function took to compute using R5R, or what network was used for the road travel times. I assume OpenstreetMap, but this needs to be stated.

## I think that the shapefile in your GitHub repo has about 100 geographic areas, which would make for 100,000 interactions from every possible pair? I would like to see more on how big the inputs are as 100x100 is very different from 1000x1000. I assume it scales with O(n^2), and that you're using R5R to compute 100,000 shortest paths? Is this Dijkstra, an optimised contracted hierarchies algorithm, CPU or a GPU algorithm? This would limit the applicability of the technique to larger sized metropolitan areas if the compute scales as n^2 and the calculation of the cost function is computationally intensive.

* Thank you for these suggestions to clarify what is a R data-package and how aspects of the R data package were calculated for those who are unfamiliar to R.
* To more clearly define a R package for new-to-R users, the Introduction Section is revised as follows:

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).

* There are 103,076 origin TAZ to destination TAZ pairs in the ‘od’ table. Inspecting this table, each origin TAZ is matched with between 1 to 241 destination TAZ. This means practically across the GGH, the total number of workers at an average origin TAZ have a total of approximately 35 work place TAZ destinations on average. There are 3,764 TAZ in the GGH, and not all TAZ have workers and/or places of employment, thus the TTS 2016 yields 103,076 origin destination pairs, so for those pairs we associate a travel time.
* ‘r5r’ is an interface for the java-based r5 routing engine (https://github.com/conveyal/r5). The outputs from r5r was a 3756\*3756 travel times, but only the travel times from existing OD pairs was retained. The aim of including the travel time within this data package is not to demonstrate the benefits of ‘r5r’, but to associate an estimated ‘cost of travel’ variable to the TTS 2016 data set in an R environment.
* Your questions on computational efficiency, and details about the ‘r5r’ method are important. We’ve added additional details about ‘r5r’ and direct readers to additional a source to see details about commutation efficiency associated with the ‘r5r’ package as follows:

Also included in {TTS2016R} are the estimated travel times between OD as summarized in descriptive statistics table in Figure 1; travel times are calculated using the package {r5r}. {r5r} interfaces with the java-based {r5} routing engine developed separately by Conveyal (Conveyal, 2022). 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 OpenStreetMap 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](<https://ipeagit.github.io/r5r/index.html>).

* We also add the following to the text discussing Figure 3 (travel times) to emphasis the uncongested nature of the road network:

Additionally, travel times (by car) within the core of the Toronto census metropolitan area (CMA) (blue) is also moderately low since traffic congestion is not reflected in the travel time calculations. Further from these areas, travel times are higher.

## I've got a few minor points as follows:

## P2, line 26: What is "n=3764 within the survey boundaries"? I couldn't understand what data you were using to make the travel-to-work flows.

* Thank you for your comment asking to clarify what data is used to make the travel-to-work flows, specifically “P2, line 26: What is "n=3764 within the survey boundaries"?”. We removed the ‘n’ terminology all together and added additional explanation regarding the significance of TAZ:

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 km2 in spatial area to a maximum of 879 km2 (median:1.3 km2 and 3rd quantile: 2.8 km2).

## P2, line 38: What is the network used for travel time calculations? (also see P.4 line 35)

* Thank you for you question asking to clarify travel time calculations (P2, line 38 and P.4 line 35). Good point. In the response to your first comment, we clarify the use of the network. We repeat the text that is added to the revised manuscript as follows:

Also included in {TTS2016R} are the estimated travel times between OD as summarized in descriptive statistics table in Figure 1; travel times are calculated using the package {r5r}. {r5r} interfaces with the java-based {r5} routing engine developed separately by Conveyal (Conveyal, 2022). 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 OpenStreetMap 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.

## P3, line 34: "and jobs is the TTS 2016 region" should be "jobs in the..."?

* Thank you for pointing out the grammatical error in P3 line 34, it has now been corrected.

## P4, line 35: What is the departure time used? Is the travel time the same across the day? Does R5R take into account rush hour traffic congestion or this a simple "20 miles driven on a 40mph limit road takes 30 minutes" calculation? I'd like to see a bit more information. Also, one mode (road) for all commuter transport seems like an over-simplification. How much rail travel is happening in the area? Walk/bike etc?

* Thank you for your questions regarding the need for clarity in the travel time calculations (P4, line 35).
* Details about the departure time were removed – since this a static OpenStreetMap road network was used, traffic conditions are neglected so time of departure is not used in the car travel time calculations. We hope we sufficiently clarified the type of network used for the travel time calculations in addressing your comment above.
* We chose to only include the travel times for car commuting because they are so critically important to the GGH. Reflecting on your comment, we added the number of work-related trips for each OD by primary mode, to the OD table in the package. With this data, the modal split was calculated and overall, 79% of trips taken using a car. We added in this context into the manuscript as follows:

Additionally, car travel is included in this data package because it is a critically important commute mode in the GGH. 2,598,379 of the trips are made using a car 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).

## P5, line 29: Where you reference Toronto and Hamilton on the choropleth maps, it took me a while to realise that I had to refer back to Fig 1 which has the names. The "south-eastern border" and "eastern areas" descriptions had me looking in the wrong areas on Fig 3, until I looked back to Fig 1. It needs something on the Fig 3 map to highlight the areas for those of us who don't know Canada. Or at the very least "refer to Fig 1 for names".

* Thank you for identifying this difficult in interpreting Figure 3 (P5, line 29). We updated Figure 3’s palette and added your suggestions in the text describing Figure 3 as follows:

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 low since traffic congestion is not reflected in the travel time estimations. Further from these areas, travel times are higher.

We hope that these modifications satisfy all the comments you communicated. Thank you greatly for your time in reviewing our paper, we believe its quality has been enhanced.

# Response to Referee #2:

## Here are a few comments that might be helpful:

## The link to access the original dataset through the ‘http://dmg.utoronto.ca/survey-boundary-files' page provided in the GitHub documentation was not working at the time of this review.

* The link (http://dmg.utoronto.ca/survey-boundary-files) seems to be working now! Thanks for flagging this, we will monitor it to ensure this link remains live.

## p. 2, l. 24-30: TAZ are central to the dataset and, as such, should be described more precisely. Could you provide an indication of their sizes and regularity? -- “Based on land-use and population demographics”, such as? and how?

* Thanks for this comment about the need to clarify TAZ (p. 2, l. 24-30). We agree TAZ are critical to the dataset. We have added an additional sentence to describe the TAZ used. The text is revised as follows:

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 km2 in spatial area to a maximum of 879 km2 (median:1.3 km2 and 3rd quantile: 2.8 km2).

## p. 3, l. 35-42: It would be interesting to indicate how the workers that are travelling outside of the GGH boundaries are identified in the data (i.e., what destination does the origin-destination table record for them).

* Thank you for this comment regarding workers traveling outside of the GGH boundaries (p.3, 1. 35-42) – we wondered about this too!
* To our knowledge, the TTS does not provide data on where workers inside the TTS travel for employment, only that they travel ‘externally’. The majority of the workers who travel externally are in closer proximity to the edge of the GGH, which makes sense.
* Additionally, the GGH accounts for approximately 55% of Ontario’s entire population and captures many of the provinces census agglomerations.
* Considering the significant of the GGH population and the limitations of the TTS, we can’t provide any additional information about the externally-travelling GGH workers.

## p. 4, l. 36: “a small proportion of trips are taken by non-car modes”: is there a way to know how much from another source?

* Great suggestion regarding the proportion of trips taken by non-car modes (p. 4, l. 36).
* We added the number of trips taken from O to D using a primary mode (car, transit, active, etc.) to the OD table so we were able to identify how prominent car mode is. As such, we replaced the text describing the car-commute in the GGH as follows:

Additionally, car travel is included in this data package because it is a critically important commute mode in the GGH. 2,598,379 of the trips are made using a car 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).

## p. 4, Figure 2. The equivalent plots in the GitHub documentation are far more readable. Would it be possible to change the palette to something else?

* Great suggestion regarding p.4 Figure 2 – we updated the figure.
* In the GitHub readme, the figures shown just use a continuous scale – this makes the map look nicer. In the Figures included in the manuscript, we used a different plotting pack and set the scale to jenks distribution to emphasis the differences in the origin manuscript submission.
* However – we changed the palette to Blue-Yellow and muted the colours so the figures in this submission are more visually appealing from our perspective but still have a scale set to a jenks distribution. We like it more now - hope you like it too!

## p. 5, l. 32: please indicate where Peterborough is located on the map.

* Thank you for pointing out that it is difficult to identify where Peterborough is located on the map (p.5, 1. 32). We’ve added a suggestion from referee #1 that clarifies this issue as follows:

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 low since traffic congestion is not reflected in the travel time estimations. Further from these areas, travel times are higher.

## p. 7, l. 7: “are included as a compliment”

* Nice catch (p. 7, l.7), we rephrased the concluding remarks sections. The parts pertaining to ‘complimenting’ data now read:

This paper introduces {TTS2016R}, an open data product in the form of a R data

package. This package is a fusion of data from multiple sources and we demonstrate

the spatial and numeric extent of the data contained within. 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.

We hope that the modifications made as suggested by your comments are to your satisfaction. Thank you for your time and insightful feedback.

# Response to Editor

## I have received two reports on your submission, and I have also inspected it myself in detail. My sense is that there is value in the manuscript and the data product, but it needs to be further "unearthed" to make it as useful as possible for the EP-B readership. Both referees provide useful comments I'd like you to address (in the case of R.1, I don't expect you to include detailed computational comparisons and this is not the venue for it, but I do think it'd be useful to try to use their comments as guidance to make your paper clearer as to what it tries and doesn't try to do). Beyond the two reports, there are two further aspects I think would be useful to incorporate in the paper:

* Thank you for these comments. We hope we addressed the comments from the two reviewers up to your satisfaction.

## Contribution of the data product: what is new? From my reading, it is not entirely clear if what you have packaged is entirely provided by TTS or you have done an additional exercise in fusing the data with additional sources (e.g., census). Could you please walk the reader through what is the data product adding above and beyond what is already available for direct download from the official TTS page? If there is only data that is already available from the official source, I'd consider combining it with other data that you think would be useful in this context.

* Agreed. Your suggestion along with the other reviews inspired us to incorporate more data into the data package. We added the number of trips, by mode, to the origin destination table. This will facilitate the analysis of modal trends in commute.
* Furthermore, we added IDs to the TAZ (which are defined by the 2016 TTS) that correspond to the Canadian Census. This will facilitate the simple integration of any census variables which users of the package can easily extract. These variables can be widely used for a variety of analysis.
* We edited the introductory section to better capture the contribution of the data product as follows:

This manuscript presents the open data product [{TTS2016R}](https://github.com/soukhova/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 @Arribas2021open]. 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 and trips by mode are sourced from the 2016 Transportation Tomorrow Survey (TTS) [@data\_management\_group\_tts\_2018], estimated travel times are calculated, and boundary files from the TTS and from the Canadian Census.

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, {TTS2016R} includes full-time home-based work-to-job origin destinations 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} [@Pereira2021r5r].

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](https://dmg.utoronto.ca/idrs/index), the raw data can be technically demanding, cumbersome to work with, and requires multiple software to process. By pre-processing the data in a `R` environment and packaging it alongside complimentary data, {TTS2016R} offers a slice of the TTS data which 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.

## Purpose and value: there is a useful example of what the package can be used for, but it is very specific. I'd like to see a more general reflection on how you expect this package to be used by the wider community (e.g., Urban Data/Code readership).

* Excellent suggestion. In addition to editing the introduction, we added a few addition remarks to the conclusion to crystallize the purpose and value of this data package to the Urban Data/Code readership.

{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.

We hope that these modifications satisfy your recommendations and the comments communicated by the reviewers. Thank you for taking the time to provide this insightful and constructive feedback!