

# ENHANCING PUBLIC TRANSPORT RESEARCH THROUGH IMPROVED DATA QUALITY AND ACCESSIBILITY: A CASE STUDY IN JOHANNESBURG

**C KARSTEN, C BEETGE and P BURGER**

Council for Scientific and Industrial Research (CSIR), PO Box 395, Pretoria, 0001

## **Conference details:**

Southern African Transport Conference (SATC)

8 – 11 July 2024

CSIR ICC, Pretoria, South Africa

## **ABSTRACT**

Public transport is used by the majority of the South African population to commute between home, school, other amenities and their place of work. There is, however, a severe deficiency in the quality of public transport data in South Africa. The absence of a standardised public transport data system is detrimental to researchers and commuters who make use of this data. This paper examines the adverse effects of inadequate public transport data on research, public transport planning and urban planning. It highlights the challenges resulting from inadequate public transport data and its detrimental effects on decision-making. Furthermore, the paper highlights the myriad of benefits that standardised, and easily accessible transport data can bring. The paper's goal is to demonstrate the critical need for enhanced public transport data in research and planning and discuss potential solutions for improving data availability. This will pave the way for a more informed and efficient approach to transportation policy and practice.

## **1. INTRODUCTION**

Public transport plays a pivotal role in the daily commuting landscape for the majority of South Africa's population, acting as the link between homes, workplaces, and other destinations. According to the 2020 National Household Travel Survey, approximately 73% of South Africans rely on public transport as their primary mode of commuting (Department of Transport, 2021). The diverse public transport system in South Africa includes Bus Rapid Transit (BRT), government-subsidised buses, rail, and minibus taxis, with minibus taxis being the predominant choice for about 62% of the population (Department of Transport, 2021).

Public transport data availability is low, hindering both commuters seeking access to services, as well as transport modellers and urban planners aiming to understand and model the transport system. The absence of a standardised methodology for data collection and storage exacerbates the issue, with varying levels of accuracy, completeness, and update intervals across different agencies. This scarcity and poor quality of public transport data in South Africa impedes effective planning at the metropolitan, provincial, and national levels. This can also impede usage from potential commuters as they do not have access to the required information.

### 1.1. Problem statement

Within the context of South Africa's public transportation system, this paper addresses the critical challenge regarding the quality and accessibility of public transport data. The absence of a standardised data capturing and storing system and a centralised repository significantly hampers researchers and urban planners, impeding their ability to make informed decisions regarding public transport.

### 1.2. Aim of paper

The paper aims to demonstrate the critical need for accurate and enhanced public transport data in research and planning and discuss potential solutions for this need in South Africa. This may result in a more informed and efficient approach to transportation policy and practice.

### 1.3. Scope of paper

This paper outlines the challenges arising from the lack of quality and accessible public transport data in South Africa. It explores existing gaps in the current public transport data landscape, emphasising the absence of a standardised data capturing and storing protocols. The potential advantages of implementing a standardised and easily accessible public transport data system are highlighted, drawing from a case study on the currently available bus data in Johannesburg. The data used includes subsidised bus and non-subsidised bus routes data and excludes minibus taxis as detailed data for minibus taxis was not publicly available at the time of the study. The paper seeks to illustrate how improved data quality can contribute to more informed and efficient approaches to transportation policy and practice. Additionally, the paper discusses potential solutions and mitigation strategies for establishing a robust and standardised data infrastructure conducive to effective decision-making in transportation policy and planning.

### 1.4. Paper structure

This paper commences with an overview of the current state of public transportation in South Africa and the benefits of quality public transport data in Section 2. Followed by the case study introduction and methodology in Section 3. In Section 4, the analysis that can be done with quality accessible public transport data is shown and the benefits of having these results are discussed. The conclusion is provided in Section 5.

## **2. LITERATURE REVIEW - OVERVIEW OF THE CURRENT STATE OF PUBLIC TRANSPORT IN SOUTH AFRICA**

As South Africa grapples with the challenges of rapid urbanisation and the pressing need for sustainable development, the integration of effective transport and urban planning becomes paramount. This literature study explores key indicators essential for addressing the unique dynamics of the South African environment, emphasising the role of public transport in fostering accessible, equitable, and environmentally conscious urban spaces. Potentials for creating high-quality public transport datasets are considered.

### 2.1. Public transport data

The National Transport Policy White Paper of 2021 (hereafter referred to as the 2021 White Paper) outlines a comprehensive vision for a transport system that is equitable, reliable, economically sustainable, and environmentally friendly (Department of Transport, 2021).

The South African government's objectives encompass supporting national development goals, ensuring basic accessibility, fostering economic growth, and engaging stakeholders in key decision-making processes. To realise these goals, the policy emphasises the importance of an accessible, cost-effective, time-efficient, reliable, safe, and secure transport system (Department of Transport, 2021).

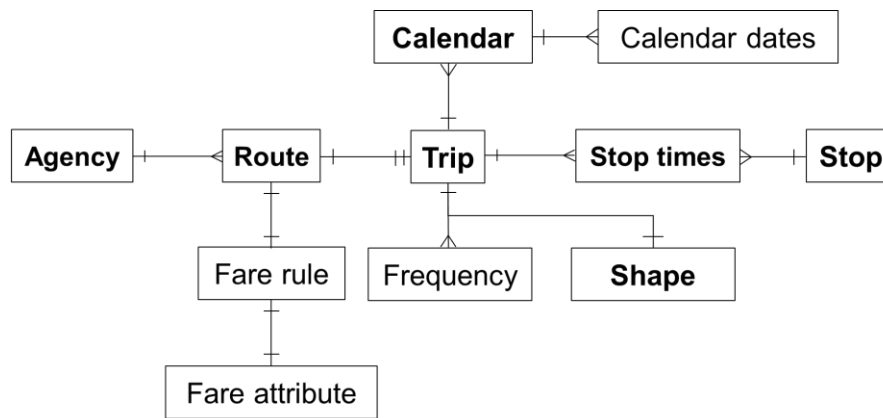
The absence of an integrated national transport database, as highlighted in the 2021 White Paper, poses a significant challenge. According to the 2021 White Paper this challenge stems from misalignments in planning, development, and investment in transport infrastructure. Adequate planning relies on accurate and standardised data collection and storage in a format that maximises utility for various stakeholders. The integration of datasets among government departments is crucial for cohesive decision-making, progress monitoring, and performance evaluation (Department of Transport, 2021).

Through work done for the Gauteng Department of Roads and Transport in the digitisation of Gauteng bus data it was seen that currently, there is no standardised methodology for data collection and storage among transport agencies. This leads to variations in accuracy, completeness, and updating intervals. The lack of uniformity in methodologies, ranging from General Transit Feed Specification (GTFS) to Microsoft Word and Excel, hinders the seamless integration of data and prevents a comprehensive understanding of the entire transport network.

The last of the nine strategic objectives outlined in the 2021 White Paper emphasises the need for a comprehensive transport data and information system (Department of Transport, 2021). GTFS is an international standard for public transportation schedules and geographic information, facilitating efficient data sharing among transport agencies, passengers, and researchers (Google Transit, 2022). High quality GTFS is up to date, accurate, and complete. This GTFS standard is beneficial for transport agencies, passengers, and transport researchers. A GTFS feed is a collection of comma separated values (CSV) files contained within a .zip file. Together, the related CSV tables describe a transit system's scheduled operations. The .zip file can contain files that each describe the following:

- Agency (information of each operator)
- Calendar (days that the route is driven)
- Calendar dates (exceptions to the standard service schedule)
- Fare rules (describes the cost structure for the trip)
- Fare attributes (monetary value linked to fare rules)
- Frequencies (headways and schedule information)
- Routes (routes of the agency)
- Trips (bidirectional variations of routes)
- Shapes (spatial coordinates of routes)
- Stop times (sequence and interarrival times between stops on a route)
- Stops (stop location information)

The bare minimum files required for a valid GTFS dataset are Agency, Calendar, Routes, Trips, Shapes, Stops and Stop times. The entity relationship diagram (ERD) of the files in a GTFS dataset is provided in Figure 1. The specification is designed to include the required information to provide trip planning functionality but is also useful for other applications such as analysis of service levels and some general performance measures. This is also the format used by Google Maps to publish public transport data (Google Transit, 2024).



**Figure 1: GTFS entity relationship diagram.**

However, challenges persist, including the fragmentation of data due to the varying collection and storage methods used by different agencies. Efforts to generate GTFS feeds are hampered by datasets being sold by the private sector or not being made publicly available. The absence of a centralised platform for all public transport feeds and the lack of maintenance for existing datasets further hinder access to accurate and up-to-date information (Gumbo & Moyo, 2020).

The collaborative efforts of government agencies, private companies, and research institutions are identified as the key to overcoming challenges and creating a more comprehensive and integrated public transport data ecosystem (Paulsson, et al., 2018). Sharing resources and expertise can address data silos and enhance the overall quality of information.

## 2.2. Public transport related decision-support

There are various ways in which reliable and high-quality transport data can assist in decision-support for urban planners or transport planners. Numerous indicators can be calculated using this data that will lead to a better integrated and more equitable society with and utilised public transport network. These include:

- **Service Accessibility:** In the South African context, evaluating the coverage of public transport services is critical to ensuring widespread accessibility. The vast and diverse geography of the country demands a comprehensive analysis of service coverage (Luke & Heyns, 2020). Identifying service gaps becomes imperative to address disparities and plan targeted expansions or modifications of public transport routes or social facilities, aligning public transport services with the diverse needs of the population (Thondoo, et al., 2020).
- **Infrastructure Utilisation:** In the South African context, where cities are characterised by unique spatial layouts, assessing the usage of different stops and stations is vital (Weber, et al., 2016). Efficient infrastructure utilisation is essential for optimising investments and ensuring that facilities align with the dynamic demands of the population. This indicator aids in crafting targeted infrastructure development strategies that cater to the specific needs of South African urban centres.
- **Affordability and Equity:** Addressing issues of affordability and equity is paramount in a country with diverse socio-economic backgrounds. Analysing fare equity ensures that public transport remains accessible to different income groups, fostering inclusivity. Simultaneously, examining service equity becomes essential to distribute

public transport service equitably, avoiding disparities and providing equal access to all communities (Hernández, 2017).

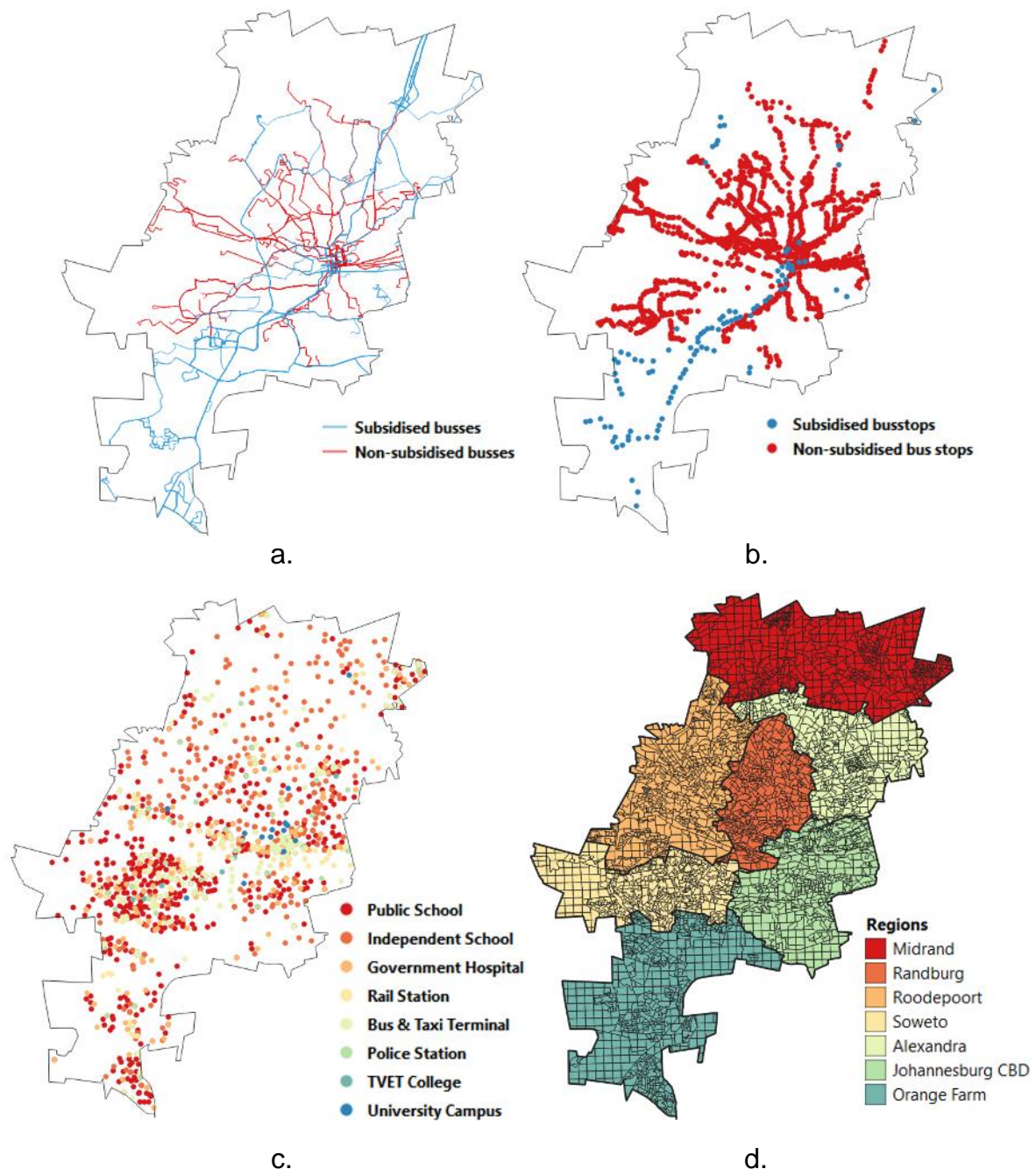
- Transit-Oriented Development (TOD) Metrics: South Africa's unique urban challenges require a focus on TOD. Assessing population and employment density around transit stations aids in supporting transit-oriented development as stated by Olaru, et al. (2011) and Uddin, et al. (2023). Additionally, examining the land use mix around transit nodes becomes crucial for creating vibrant and mixed-use urban environments that align with South Africa's cultural and economic diversity (Mushongahande, et al., 2014).

In conclusion, this literature review highlights the crucial need for high-quality and accessible public transport data in South Africa's evolving urban landscape. The absence of a standardised national transport database impedes comprehensive decision-making, emphasising the urgency for data standardisation. While the potential of GTFS is acknowledged, persistent challenges call for collaborative efforts to ensure data accuracy and accessibility. The review highlights the pivotal role of reliable transport data in shaping informed decisions for urban planners, emphasising indicators such as service accessibility, infrastructure utilisation, affordability, equity, and TOD metrics. Prioritising the acquisition and collaborative enhancement of high-quality public transport data is fundamental for fostering inclusive and sustainable urban spaces in South Africa.

### 3. CASE STUDY AND METHODOLOGY

#### 3.1. Case study and study area

To demonstrate the impact of having public transport data available in GTFS format, a case study was conducted in Johannesburg. GTFS datasets for the non-subsidised (metrobus, BRT), and some of the government-subsidised buses were used to analyse the decision-support that can be provided to urban and transport planners with public transport data. The coverage of these public transport routes is shown in Figure 2**Error! Reference source not found.a** and the bus stops are shown in Figure 2**Error! Reference source not found.b**. This case study demonstrates use of subsidised and non-subsidised bus routes data in various accessibility analyses to points of interest (POI)s and decision-support in densification and network expansion. The POIs are mapped in Figure 2**Error! Reference source not found.c**. The POI locations were partly derived from extracting certain land use classes from the building-based land use dataset and was partly derived from datasets obtained from the City of Johannesburg describing health, education, and other facilities. Household and job projections for 2024 for Johannesburg on zonal level was obtained from the urban growth simulation model UrbanSim (Waddell, 2002), adapted and implemented by the Council for Scientific and Industrial Research (CSIR). Since the focus of this paper is not on land use modelling, the interested reader is referred to Waldeck, Holloway and van Heerden (2020), who provide an overview of the urban growth model.



**Figure 2: Spatial visualisation of data used in the analysis for the study area. a. Bus routes b. Bus stops c. Points of Interest d. Zones and Regions**

### 3.2. Public Transport OD Matrix

To calculate the accessibility, various origin-destination (OD) matrices were required. The entire Johannesburg was divided into analysis zones of approximately 1 km<sup>2</sup> as shown in Figure 2Error! Reference source not found.d. Using these zones, a matrix was created with the BRT and metro bus GTFS datasets. In addition, an OD matrix was created for the subsidised buses. Finally, an OD matrix was created using all the above public transport routes.

To generate the OD matrix a list of trips is required with their associated distances and costs. To generate this list for an agency the trips, shapes, stops, and stop times in the GTFS feed are used. For each trip the relevant stops, stop times, and shapes are obtained. Starting at the first stop the distance to all the other stops are calculated using the shape file associated

with the trip. The index of the nearest coordinate of the shape is found for the starting and ending stop. The distance between the stops is then estimated by summing the haversine distances between the starting and ending coordinate indices. This process is repeated for all the stops associated with the trip. The only difference is that stops later in the trip do not have access to preceding stops. Therefore, the distances for the stops that follow are calculated. This distance information is combined with information about the fee structure of the agency. For example, if the agency uses distance-based costing, it can easily be determined from the calculated distances. Some agencies also make use of flat rates. Following this process, the cost and distances have been calculated for each stop combination that a commuter has available to them. These are termed stop pairs and define the stops reachable and the associated cost for a given starting bus stop.

To generate the public transport OD matrix the stop pairs are used in conjunction with the zonal centroids. To generate a single entry in the OD matrix the origin and destination zone centroid coordinates are obtained. Using the centroids of the zones, the available stops can be determined by calculating the haversine distance between the stops and the centroids and defining a maximum cut-off distance. This distance determines which stops the zones have access to. Now that the available stops for the origin and destination zones have been determined, the list of stop pairs can be filtered for entries that match both the origin and destination stops specified for the zones. Depending on the type of analysis the shortest distance or lowest cost option can be determined from the filtered list of stop pairs. The distance and cost are recorded for the origin and destination combination. This process is repeated for every possible combination of origin and destination zone.

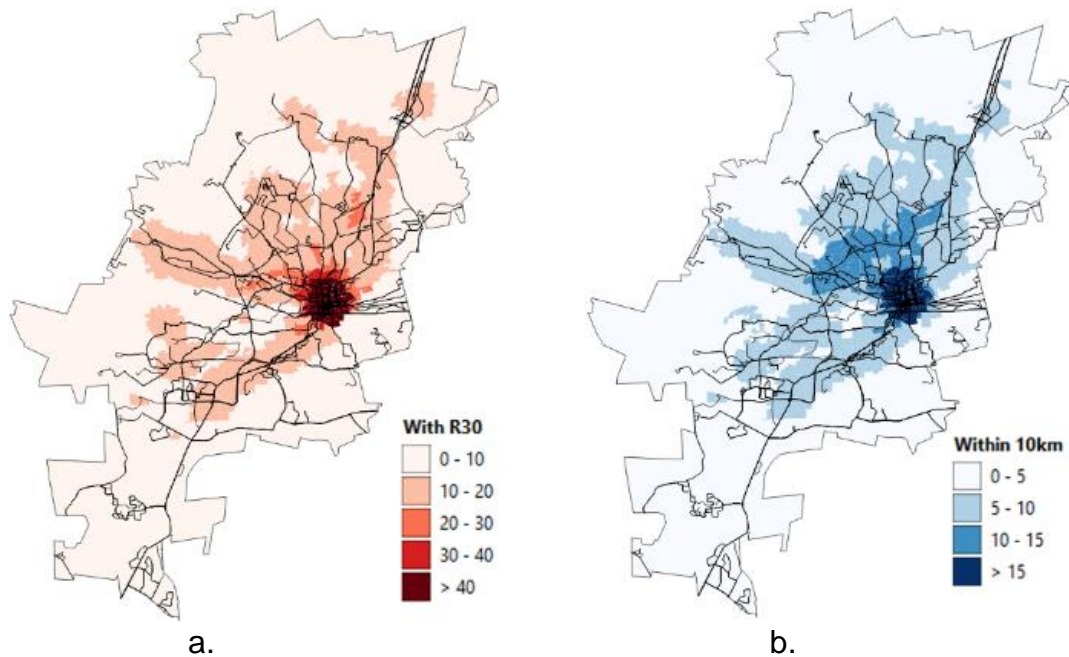
## **4. RESULTS**

Efforts were made to collect as much of the GTFS data for Johannesburg as was feasible. As stated previously, the data is still incomplete due to missing bus agencies and other transport modes. The completeness of the data is estimated at around 70% of the scheduled public busses. Therefore, the data is still sufficient to allow for initial investigations into the accessibility of the study area and to demonstrate the analysis methods that become available when GTFS data becomes available. As more GTFS data is collected and more agencies make use of the format and make their data available to researchers and planners, the accuracy of the results can be further improved.

### **4.1. Overall Accessibility**

Accessibility can be investigated from numerous different perspectives. For example, by placing cost and distance restrictions on the generated OD pairs the resulting indicators can differ substantially. Restrictions filter the OD matrix for values that fall below a certain threshold. The filtered list of OD pairs can then more accurately represent how the public engages with public transport. For example, a trip can exist to a POI, but due to the travel time or cost associated with it, it becomes infeasible to the average commuter. Figure 3 shows the percentage of Johannesburg's area that is reachable using public transport with different trip restrictions. The figure is generated by determining the zones that can be reached from each zone (using the OD pairs) and summing the reachable zone areas and dividing by the total area of Johannesburg. The figure indicates that for 30 Rand, more than 40% of the city's area is accessible from the central business district (CBD) when using a scheduled subsidised or non-subsidised bus service. As stated in Section 1.3, minibus taxi routes are excluded from the analysis. The accessibility decreases substantially when looking at the percentage of Johannesburg reachable when trip distances are restricted to 10km. It can also be seen in Figure 3 that accessibility is generally higher for zones that are

near bus routes and stops.



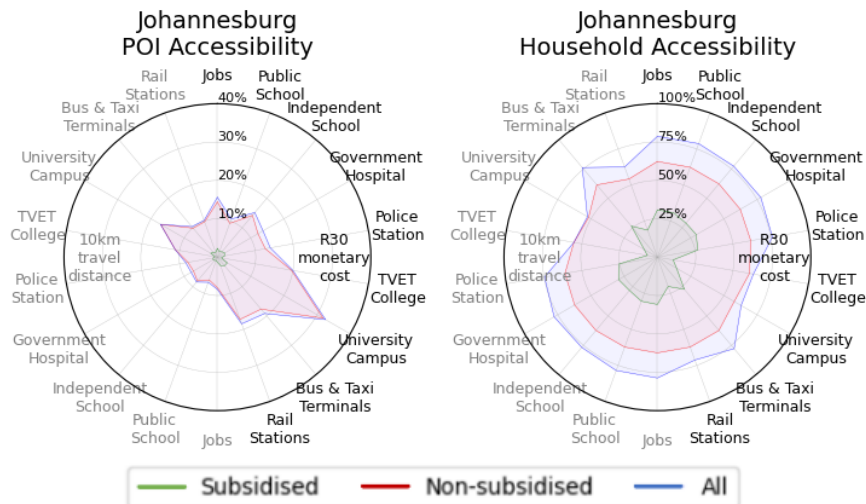
**Figure 3: Percentage of the area of the total Johannesburg accessible for each zone with scheduled bus transport routes. a. Within R30 cost threshold b. Within 10km distance threshold**

The National Public Transport Subsidy Policy Draft of 2023 states that the income spent on public transport should be limited to 10% of a person's income (SABOA, 2023). To put the R30 limit into perspective, this amounts to a monthly salary of around R12 000,00. This accessibility will decrease when a lower monthly income is used. The R30 limit was used to demonstrate how accessibility can be calculated within a monetary context.

Determining what percentage of the city is reachable with public transport does not create a complete picture of accessibility. Commuters are generally trying to access specific POIs such as schools and hospitals with public transport. For this reason, the access to these POIs needs to be determined. One way to do this is to calculate the percentage of each POI type that is accessible from each zone. This is done by taking the set of zones reachable by a zone and summing the POIs in the reachable zones and dividing this by the total number of POIs for the POI type.

Another measure is percentage of households that have access to at least one POI. This is done by summing the number of households per zone that can reach at least one zone that contains the relevant POI and dividing it by the total number of households in all zones. Again, the OD pairs are reduced by applying a cost and distance threshold respectively. Figure 4 shows the results of applying these techniques to the study area. From the figure it can be seen that the average zone has access to about 10% of the POIs for each POI type using public transport. Again, the distance constraint proves to be more restrictive than the cost constraint. The figure also shows that approximately 75% of all the households in the study area have access to at least one of the POIs for each of the POI types, if they are able to spend R30 per trip on these scheduled bus services. It is also clear that the non-subsidised buses offer greater access to POIs in comparison to the subsidised buses. This could, however, be due to data availability. GTFS data of more agencies and different

transport modes can significantly increase these percentages as the network reach will improve.

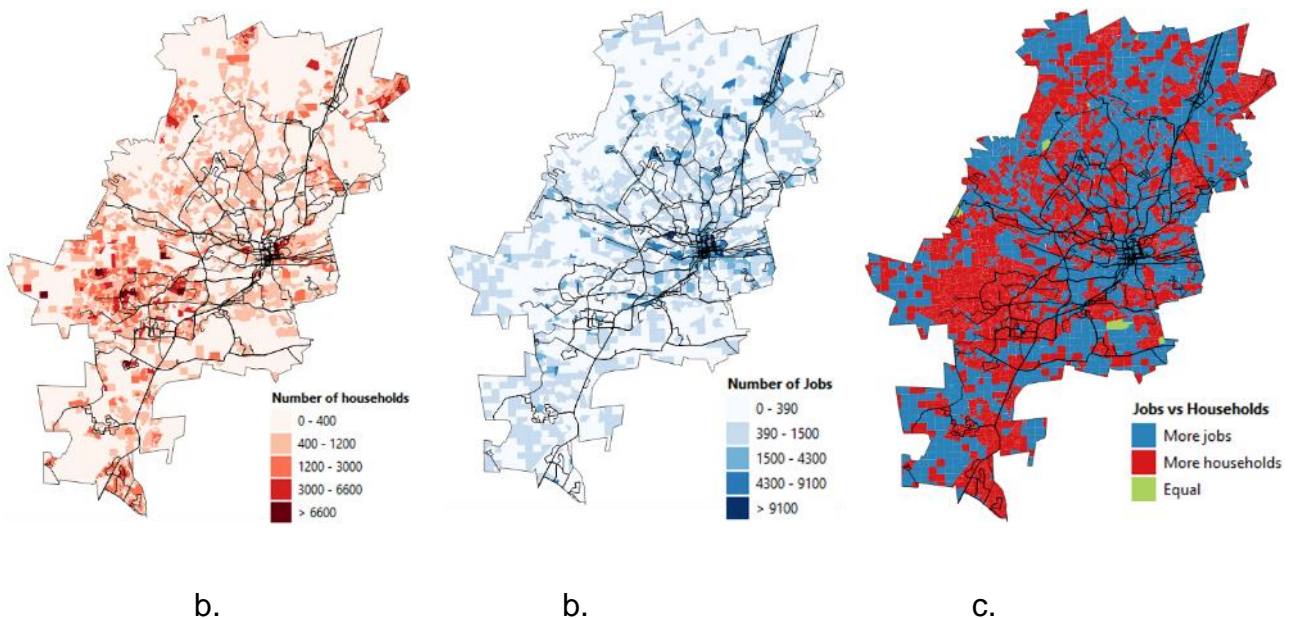


**Figure 4: Overall accessibility measures for Johannesburg at a fixed cost or travel distance for different bus service types. Left: Average percentage of points of interest accessible for each zone. Right: Percentage of zones that have access to at least one POI.**

#### 4.2. Accessibility Equity

Public transportation stands as the predominant mode of transport for the majority of South Africans, particularly among low-income households. Access to high-quality public transport data proves instrumental for urban and transportation planners in ensuring that existing networks effectively connect these low-income households to various POIs and employment opportunities. Figure 5 provides a concrete example of this relationship. While the public transport network examined in this study (scheduled subsidised and non-subsidised bus routes) does extend its reach to regions with lower-income households, a substantial area remains underserved, as depicted in Figure 5. This observation serves as a crucial foundation for identifying strategic locations for network extensions. Additionally, it aids in determining the most beneficial type of service, considering factors such as cost and average household income. This nuanced analysis, facilitated by quality public transport data, holds the key to creating a more inclusive transportation infrastructure that addresses the diverse needs of communities and promotes social equity.





**Figure 6: Distribution and land use mix with transport network for Johannesburg a. Household distribution b. Job distribution c. Land use mix.**

## 5. DISCUSSION

These measures discussed above are only a few that can be calculated using GTFS data that speaks to the indicators that are important for urban and transport planners. Accessibility over different time periods in a day can also be analysed using the stop times in the GTFS dataset. Through these examples, the importance of high quality and accessible public transport data is highlighted.

There are various other alternatives to GTFS such as Service Interface for Real Time Information (SIRI), Linked Open Data (LOD) that uses Resource Description Framework (RDF). However, GTFS stands out as a preferred format for public transport data due to its standardised structure, facilitating interoperability and accessibility across different transit systems. Its adoption fosters innovation and ease of integration with mapping services, enhancing urban mobility solutions. Despite challenges like complexity in maintenance, GTFS remains popular for its widespread usage, cost-effectiveness, and alignment with industry standards. Its pivotal role in efficient data sharing underscores its continued preference among transit agencies and stakeholders.

Although there are various formats in which public transport data can be stored, GTFS is by far the most widely used format (Ngoma, 2021). This is also the format used by Google to publish transport data on Google Maps. The GTFS format provides high quality data that can be used in various instances. Once the static GTFS data is available, it can be enhanced to provide commuters accurate information in the form of a real time display of where the transport currently is and available trips and stops and stop times. Google Maps have the option of planning a trip using the available GTFS data. It also provides transport and urban planners the required information for better accessibility and connectivity planning. As mentioned in Section 2.1, transport data is stored in various data formats including Microsoft Word and Excel, which is not suitable since it impedes integration with other datasets, and it is difficult to maintain.

Various options for storing and sharing of GTFS data are available. The CSIR has, in recent years, developed a conceptual model of how a centralised, standardised, public transport database could assist in improving data quality and availability. In addition, a GTFS Platform

was subsequently developed to solve many of the problems discussed in this paper. This platform is currently being modified for external use. The platform aims to assist users to easily create their own GTFS feeds in the correct format. This GTFS Platform can be used to upload, create, maintain, manage, and share public transport data in GFTS format in the form of GTFS datasets. These GTFS datasets can in turn be used in transport modelling models or uploaded onto Google Maps or similar apps to make the information more accessible to the commuters.

## 6. CONCLUSION

The absence of an integrated national transport database in South Africa, as emphasised in the 2021 White Paper, presents a substantial hurdle, causing misalignments in planning and development. This review highlights the critical role of accurate and standardised data, essential for effective planning and aligned decision-making across government departments. Examining the state of public transport data in South Africa revealed a lack of uniformity in collection and storage practices among agencies. To address this, GTFS was recommended as a standardised format, ensuring high-quality data for various analyses. A case study in the Johannesburg metropolitan area illustrated the decision-support capabilities derived from accessible quality transport data. For data collection and management, the CSIR's GTFS Platform emerged as a recommended solution, offering the ability to create, store, and share GTFS data publicly or privately. The data can also be made available on Google Maps to make it accessible to commuters. This approach highlights the significance of establishing standardised, accessible, and high-quality transport data systems to advance efficient and informed transport and development decision-making in South Africa.

Future work on this topic could aim at including the data that could not be obtained for this analysis. The minibus taxi routes cover a substantial part of the study area and are priced comparably to the bus operators. Including this data could have profound effects on the results. In addition to this, the analysis could be expanded to allow for multi-modal transport where commuters can make use of more than one service to reach their destination. However, many of these trips would become prohibitively expensive and thus not meet the cost criteria set.

## 7. REFERENCES

- Department of Transport, 2021. *White paper on national transport policy*, Pretoria: Government Gazette.
- Google Transit, 2022. *General Transit Feed Specification*. [Online] Available at: <https://developers.google.com/transit/gtfs/reference>. [Accessed 15 01 2024].
- Google Transit, 2024. *Google Transit*. [Online] Available at: <https://developers.google.com/transit/gtfs> [Accessed 09 01 2024].
- Gumbo, T. & Moyo, T., 2020. Exploring the interoperability of public transport systems for sustainable mobility in developing cities: Lessons from Johannesburg Metropolitan City, South Africa. *Sustainability*, pp. 12(15), 5875.
- Hernández, D., 2017. Public transport, well-being and inequality: coverage and affordability in the city of Montevideo. *CEPAL Review*.
- Luke, R. & Heyns, G. J., 2020. An analysis of the quality of public transport in Johannesburg, South Africa using an adapted SERVQUAL model. *Transportation Research Procedia*, Volume 48, pp. 3562-3576.

- Mushongahande, R., Cloete, C. E. & Center, C. J., 2014. Impact of the Gautrain on property development around station precincts. *Journal of the South African Institution of Civil Engineering*, 56(1), pp. 2-10.
- Ngoma, L., 2021. *Public transport: The main data exchange formats*. [Online] Available at: <https://m2050.media/en/public-transport-the-main-data-exchange-formats/>
- Olaru, D., Smith, B. & Taplin, J. H., 2011. Residential location and transit-oriented development in a new rail corridor. *Transportation Research Part A: Policy and Practice*, 45(3), pp. 219-237.
- Paulsson, A. et al., 2018. Collaboration in public transport planning – Why, how and what?. *Research in Transportation Economics*, Volume 69, pp. 377-358.
- S Siuhi, J. M., 2016. Opportunities and challenges of smart mobile applications in transportation. *Journal of traffic and transportation engineering*, pp. 582-592.
- SABOA, 2023. *NATIONAL PUBLIC TRANSPORT SUBSIDY POLICY*. South Africa, Department of Transport.
- Thondoo, M., Marquet, O., Marquez, S. & Nieuwenhuijsen, M. J., 2020. Small cities, big needs: Urban transport planning in cities of developing countries. *Journal of Transport & Health*, Volume 19.
- Uddin, A. et al., 2023. A framework to measure transit-oriented development around transit nodes: Case study of a mass rapid transit system in Dhaka, Bangladesh. *Plos One*, 18(1)(e0280275).
- Waldeck, L., Van Heerden, Q. & Holloway, J., 2020. Integrated land use and transportation modelling and planning: A South African journey. *Journal of Transport and Land Use*, Volume 13(1), pp. 227-254.
- Weber, R., Tammi, I., Anderson, T. & Wang, S., 2016. A Spatial Analysis of City-Regions: Urban Form & Service Accessibility. *Nordregio Working Paper*, Volume 2.