

van der Merwe, Jacomien; Krygsman, Stephan

Working Paper

The relationship between transport accessibility and employment duration

WIDER Working Paper, No. 2020/56

Provided in Cooperation with:

United Nations University (UNU), World Institute for Development Economics Research (WIDER)

Suggested Citation: van der Merwe, Jacomien; Krygsman, Stephan (2020) : The relationship between transport accessibility and employment duration, WIDER Working Paper, No. 2020/56, ISBN 978-92-9256-813-9, The United Nations University World Institute for Development Economics Research (UNU-WIDER), Helsinki,
<https://doi.org/10.35188/UNU-WIDER/2020/813-9>

This Version is available at:

<https://hdl.handle.net/10419/229280>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

WIDER Working Paper 2020/56

The relationship between transport accessibility and employment duration

Jacomien van der Merwe¹ and Stephan Krygsman²

May 2020

Abstract: The purpose of this research is to investigate whether transport accessibility influences the employment duration of individuals in South Africa. The South African Revenue Service's IRP5 administration datasets, which indicate employment duration and spatial location data (where workers reside and work) for all income earners in South Africa, were used to determine individuals' employment duration, as well as the travel distances between their residence and place of employment. Airline distance was used as a proxy for transport accessibility. The results indicate a negative relationship between increased commuting distance and employment duration for lower-income individuals, with these commuters being more affected by greater travel distances than higher-income groups. Spatial mapping indicates a job–housing mismatch in South African metropolitan areas and limited employment search areas for lower-income workers compared with higher-income workers, leading to further inequality.

Keywords: employment duration, equality, job–housing mismatch, transport accessibility

JEL classification: J64, R12, R40

Acknowledgements: This working paper is part of Work Stream 4, 'Turning the tide on inequality', within the project 'Southern Africa – Towards Inclusive Economic Development' (SA-TIED), and is supported by SA-TIED. The authors would like to acknowledge helpful comments by participants at the SA-TIED workshop held in Pretoria in November 2019.

The authors greatly appreciate the guidance of Professor Hannelie Nel on statistical analysis, as well as the assistance and input from Professor Tom de Jong on the calculation of airline distances and the use of Flowmap.

¹ Stellenbosch University, Stellenbosch, South Africa, corresponding author: jacomienvdm@sun.ac.za; ² Stellenbosch University, Stellenbosch, South Africa, skrygsman@sun.ac.za

This study is published within the UNU-WIDER project [Southern Africa – Towards Inclusive Economic Development \(SA-TIED\)](#).

Copyright © UNU-WIDER 2020

Information and requests: publications@wider.unu.edu

ISSN 1798-7237 ISBN 978-92-9256-813-9

<https://doi.org/10.35188/UNU-WIDER/2020/813-9>

Typescript prepared by Luke Finley.

The United Nations University World Institute for Development Economics Research provides economic analysis and policy advice with the aim of promoting sustainable and equitable development. The Institute began operations in 1985 in Helsinki, Finland, as the first research and training centre of the United Nations University. Today it is a unique blend of think tank, research institute, and UN agency—providing a range of services from policy advice to governments as well as freely available original research.

The Institute is funded through income from an endowment fund with additional contributions to its work programme from Finland, Sweden, and the United Kingdom as well as earmarked contributions for specific projects from a variety of donors.

Katajanokanlaituri 6 B, 00160 Helsinki, Finland

The views expressed in this paper are those of the author(s), and do not necessarily reflect the views of the Institute or the United Nations University, nor the programme/project donors.

1 Introduction

Transport is associated with a negative utility. It represents friction between where people are and where they want to be. This friction is often measured in terms of travel costs and travel time, and possibly convenience, comfort, and safety of travel. Individuals will always attempt to minimize this disutility while attempting to maximize access to employment, social facilities, education, etc. Several researchers have explored this relationship between access and employment status, income, and increased productivity. Labour force participation decisions are often based on comparing the costs of working, which include transport costs, against the wages earned from a job (Venables 2016). Any improvement in transport that lowers the generalized costs of transport (time, money, and convenience) may therefore increase labour participation and support economic growth. This is often one of the main arguments in support of transport infrastructure investment policies.

South Africa faces significant employment challenges, as reflected by recent employment statistics. Official unemployment has reached 30 per cent, although this figure may drastically underestimate unemployment among the youth and low-schooled populations. There was a 31 per cent increase in the number of discouraged workers¹ in South Africa between the first quarter of 2017 and the same quarter in 2019 (Stats SA 2017b, 2019a).

The National Household Travel Survey (NHTS) of 2013 revealed that 75 per cent of households that earn up to ZAR3,000 per month spend 20 per cent or more of their income on transport to commute to work (Stats SA 2017a). This is more than double the national benchmark of limiting transport expenditure to 10 per cent (Department of Transport 2015). This survey also indicated disproportionate expenditure on transport between low- and high-income households. High-income households spend 10 per cent or less of their household income on transport. This contributes to the income inequality in South Africa and may contribute to the high unemployment rates of lower-income groups in the country. The most recent Quarterly Labour Force Survey (QLFS), in the fourth quarter of 2019, reveals that the biggest reasons why people are not looking for work are that 69 per cent indicate that no jobs are available in their area, secondly 6 per cent indicated losing hope of finding a job, and thirdly 4 per cent indicate that they lack money to pay for transport to look for work (Stats SA 2019c).

South African cities are characterized by an unbalanced spatial structure, with lower-income and often vulnerable communities located on the periphery, far away from employment opportunities and other amenities (Sinclair-Smith and Turok 2012). The result is long travel times, and potential high transport costs to access employment and other activity engagement opportunities. This may lead to lower levels of employment and shorter employment duration for these income groups.

Investment in transport and improved accessibility can lead to direct economic benefits, such as lower transport costs and better time utilization. Improved accessibility can also lead to indirect or third-order benefits that create economic development through higher rates of employment, technological innovation, spatial agglomeration,² and increased productivity. Employment can be further unpacked by investigating the benefit of increased employment duration or lower employee

¹ Discouraged jobseekers are unemployed individuals who want to work, but feel discouraged from actively find work.

² Spatial agglomeration is the spatial clustering of industries, firms, or labour.

churn, which increases employer productivity, decreases recruitment costs, and increases employee morale.

This paper aims to establish whether transport accessibility, measured by spatial separation, has an impact on the employment duration of workers, specifically for individuals in lower-income groups.

2 Literature review

2.1 Spatial and modal mismatch

Kain (1968) investigated housing market discrimination towards African American workers in the USA and the impact thereof on unemployment. This research led to the spatial mismatch hypothesis, which states that low-income households that live far from employment opportunities will experience higher rates of unemployment. Later studies built on Kain's spatial hypothesis and highlighted that poor access to proper public transport modes, in addition to travel distance, could also have a major impact on job accessibility. This is specifically true for captive-mode³ low-income commuters with no access to private transport (Grengs 2010).

Kain's hypothesis was investigated in the South African context by Rospabe and Selod (2006). They studied the spatial segregation of where people live and work in the City of Cape Town using the Regional Services Councils' (RSC) levy database⁴ and the Migration and Settlement in the Cape Metropolitan Area 1998 dataset.⁵ They investigated the impact of spatial mismatch (between home and work location) on unemployment. A logistic regression model estimated the probability of an individual being unemployed considering various individual, household, and neighbourhood characteristics. Average commuting distance was included as an independent variable, and the researchers found that an increase in commuting distance will increase an individual's probability of being unemployed.

Blumenberg and Pierce (2014) used a panel dataset comprising data collected in five major metropolitan areas in the USA⁶ to estimate the relationship between automobile ownership and transit access, with employment retention as the dependent variable. They also used a multinomial logistic regression model and found that households which gained access to a private vehicle to commute to work increased their probability of finding employment by a factor of two. Their research also indicated an increase in employment when households moved to a neighbourhood with better public transport access.

A study conducted in the UK by Johnson et al. (2017) explored the relationship between transport accessibility (public transport travel times) and employment using cross-sectional data. Their research attempted to control the potential causal relationship between transport accessibility,

³ Captive-mode users do not possess a driver's licence or do not own a car and thus have no other choice than to use either non-motorized transport or public transport.

⁴ The RSC levy is a tax on firms that operate in the City of Cape Town and was used to create an administrative database of such firms.

⁵ The sociology department of Stellenbosch University and the Cape Metropolitan Council undertook this survey in 1998 to support spatial development in the Cape Metropolitan Area.

⁶ Baltimore, Boston, Chicago, Los Angeles, and New York.

vehicle ownership, and employment by using an instrumental variable approach. Elasticity was calculated, which indicated that a reduction in public transport travel time of 10 per cent would increase employment by between 0.13 and 0.3 per cent.⁷

Lau Cho-Yam (2010) elaborated on the history of spatial planning in Hong Kong and painted a very similar picture to the one in South Africa: the segregation of poor residents in neighbourhoods far from job opportunities, a movement from a manufacturing- to a service-driven economy, a shrinking job pool for lower-skilled individuals, and high income inequality. Lau Cho-Yam's (2010) study attempted to identify the relationship between urban decline, transport provision, and job-seeking range. It used the space-time activity approach to determine the size of the daily potential path area for each household as a proxy for access to potential job opportunities. Past and current spatial planning has led to an increasing income gap between the working poor and the high-income population. As a recommendation, the study stressed the fact that potential social impacts should be included in cost-benefit appraisals to improve transport access and increase the job-seeking range of low-income residents.

2.2 Transport accessibility and job search

Phillips (2014) conducted a study in the USA proving that unemployed individuals who receive a direct transport subsidy, which reduces their transport costs when searching for work, significantly increase their job search intensity, ultimately leading to shorter periods of unemployment. Research by Franklin (2015) supported this relationship between transport costs and job-seeking efforts by conducting a controlled group survey in Ethiopia. He provided transport subsidies to individuals who searched for work, and found that more people searched for work when they received a transport subsidy compared with the group that did not receive a subsidy.

The increase in discouraged work seekers in South Africa has fuelled the discussion on whether transport costs have an impact on job search intensity. Some calls have been made to compensate unemployed work seekers for their transport costs in order to reduce their search costs.

Data from the South African National Income Dynamics Survey (NIDS) were used by Ngarachu et al. (2015) to investigate the relationship between wages and transport costs. The research concluded that transport costs are a significant driver of inequality in South Africa and that lower-income workers find it difficult to 'transfer' the cost of commuting to their employer and to allocate a significant percentage of their salaries to transport compared with higher-earning individuals.

The impact of transport affordability on equality in South Africa was confirmed in a study conducted by Kerr (2015). He compared the commuting times for commuters of different income groups using different transport modes. The study indicated that commuters who make use of public transport—specifically bus and minibus taxi—spent a significantly higher percentage of their income on transport than private vehicle users. He also compared different datasets, such as the NIDS, NHTS, and the Project for Statistics on Living Standards and Development (PSLSD)⁸ to show the ever-increasing travel times for commuters in South Africa and the significant difference in travel times from commuters of other countries. Kerr (2015) emphasized the

⁷ 0.13 per cent for rural areas and 0.3 per cent for more dense urban areas.

⁸ Conducted by the Southern Africa Labour and Development Research Unit, SALDRU, at University of Cape Town.

importance of government spending on public transport to reduce the cost of commuting and to reduce the travel times for public transport users to benefit low-income residents.

Kgwedi and Krygsman (2019) illustrated not only that low-income workers in South Africa face long transport travel times, but that these commutes involve significant transfers, as well as access to and egress from the public transport system. The out-of-vehicle time component often exceeds the time spent on the public transport mode, which results in onerous commutes (Kgwedi and Krygsman 2019).

2.3 Employee churn

Kerr (2017) defined churning as the difference between worker flows and job flows,⁹ and found that workers change jobs much more relative to job openings at firms, which gives rise to increased churning. If worker flows exceed job flows, this indicates employee churn. Higher employee churn is associated with higher recruitment and training costs, as well as lower productivity among workers. Studies conducted to calculate worker flow and churn in companies use different datasets, including household censuses (Bassanini and Garnero 2012) and quarterly employment surveys that collect data on how firms in South Africa create or terminate jobs (Kerr et al. 2013), as well as a newly available dataset from the South African Revenue Service (SARS).¹⁰ These records indicate how workers move in and out of employment at firm level from one tax year to the next (Kerr 2017), and that more workers move in and out of employment as needed by firms to adjust to their company size.

Churning in South African firms equalled an average of 30.78 per cent, indicating that almost one-third of employees move between employment or out of employment per year (Kerr 2017). Kerr (2017) also suggested that employee churn is not industry-specific, but that churn rates are firm-specific and that human resource policies adopted by firms could be the main driver of higher churn rates.

This literature review indicates the current limited understanding of the relationship between transport accessibility and employment duration, with limited disaggregate data and spatial information that has hindered past research attempts.

3 Limitations to existing South African transport and employment data

The only source of travel data at household level for South Africa is the 2013 NHTS (Stats SA). A total of 51,341 dwelling units (households) were interviewed, resulting in travel information obtained from 157,273 individuals. The employment-related questions included employment status, location data at the enumerator area (EA) level of where the person resided,¹¹ location data on the transport zone level of where the person was travelling to, gross income, and trip details, which included the following:

⁹ Worker flows are defined by Kerr (2017) as the sum of all hires and separations in a firm in a tax year. Job flows are defined as the difference between employment in a tax year and employment in the previous tax year.

¹⁰ The SARS datasets include data on workers who earn more than ZAR2,000 per annum in South Africa.

¹¹ An EA is the smallest geographical unit (piece of land) into which a country is divided for enumeration purposes. EAs contain between 100 and 250 households.

- trip start and end time (used to calculate travel time between origin and destination);
- mode(s) used;
- trip costs;
- whether the individual received an employer travel subsidization.

The sample of 157,273 individuals who were interviewed consisted of children, workers, non-workers, and people of 65 and older. Only 22,593 individuals who commuted to (either informal or formal) work provided origin and destination locations. While the NHTS survey data is useful to analyse general transport trends and general levels of access, it is less suitable to model accessibility and is not able to determine employee churn or duration.

As the travel time information in the NHTS is stated and not actual measured travel time, it may also not accurately reflect actual accessibility.

3.1 National Income Dynamic Survey (NIDS)

The NIDS¹² interviewed approximately 30,000 individuals from 7,000 households in South Africa in five waves over a nine-year period from 2008 to 2017. One of the purposes of the survey was to track changes in individual and household income and expenditure and individual employment status.

The NIDS is useful for tracking changes in individuals' employment status and to establish whether there are income and expenditure or household characteristics that may influence the employment duration of individuals. Unfortunately, only Wave 2 of the survey included travel-related questions as part of the employment section. Information on the location where the individual worked was excluded. Other travel information collected included mode of transport, trip costs, travel distance, and travel time. While it is possible to investigate the relationship between employment duration and transport accessibility using Wave 2 of the NIDS, the sample size is not ideal.

Recently, a unique opportunity arose to use the SARS datasets to establish whether such a relationship exists, which may allow for a more in-depth examination of this relationship.

4 Data sources

The data for this paper are from the SARS datasets, which were made available through collaboration between UNU-WIDER, National Treasury South Africa, and SARS. This was supplemented with the spatial data layers of the 2013 NHTS and 2011 Census, which were used to develop distance matrices that were merged with the SARS datasets.

4.1 Distance and travel time matrices

ArcMap, together with Flowmap,¹³ which is a software package used to visualize and analyse geographical and flow data, was used to develop an origin-destination (O-D) matrix to measure airline distance (in kilometres) between all main places in South Africa. Origins and destinations

¹² The NIDS user manual can be downloaded from the DataFirst website (see Brophy et al. 2018).

¹³ <http://flowmap.geo.uu.nl>.

are represented by the centroids of each main place. This dataset consists of 197,093,521 O-D pair observations.

‘Internal trips’ refers to instances where the workplace and residence fall in the same main place. Three equations (see Table 1) were used to generate three internal trip distance measurements. The three equations differ with regard to (1) how the residences and workplaces are distributed within the main place polygon, (2) the form of the main place polygon, and (3) the perimeter and assumed average speed of the modes used for travelling. An area-based approach was used whereby the internal trip distance was derived from the size of the main place.

Research conducted by Melhorado et al. (2016) provides a more detailed analysis of the different methodologies for calculating internal trips. These three methods of calculating internal trip distances were chosen due to data availability.

Table 1: Equations to calculate internal trip distances

Internal trip distance	Equation
Internal trip distance method 1	$0.91 * \sqrt{\frac{Area}{\pi}}$
Internal trip distance method 2	$2 * \sqrt{Area} / 3$
Internal trip distance method 3	$\left(1 + \ln\left(\frac{\frac{perimeter}{2\pi}}{\sqrt{\frac{Area}{\pi}}}\right)\right) * \frac{\sqrt{\frac{Area}{\pi}}}{\sqrt{2}} / speed$

Notes: area and perimeter are the spatial properties of the main place and speed is a function of mode. The speed is assumed at 20 km/h, which is the weighted average speed derived from the Cape Town Transport Demand Model over all modes in the morning peak period.

Source: authors’ construction based on Melhorado et al. (2016).

The average internal distances and standard deviations by metropolitan area of the three methods were compared with the same O-D distances recorded in the NHTS 2013. Method 2 was considered the best fit and was thus used in this study. This equation assumes that employment opportunities are more concentrated within the main places compared with the concentration of households.

The airline distances, together with the internal trip calculations, were used as a proxy for accessibility in this study.

4.2 South African Revenue Service (SARS) data

South African employers that are registered for pay-as-you-earn (PAYE¹⁴) are required to provide their employees, irrespective of the amount of remuneration, with an IRP5 certificate (employee tax certificate). These certificates contain information about the employer, such as the company

¹⁴ PAYE is a system where the employer deducts income tax from the employee’s salary and pays it directly to the revenue service—SARS, in South Africa’s case.

name, physical address, PAYE registration number, and their Skills Development Levy and Unemployment Insurance Fund information.

Information regarding the employee includes their name, physical address (home), bank details, period of employment, and income information such as total income, allowances, and fringe benefits received for the tax year. It also reflects any deductions made regarding pension fund and medical aid.

The IRP5 datasets for the period 2011–17 were made available to the researchers. These datasets were already anonymized whereby no specific person or entity could be identified through the data. Additional fields attached to the individual IRP5 certificates include gender, employee's date of birth, and a main place identifier. The main place identifiers do not indicate any home address. The data were cleaned by us; this cleaning is described in detail below.

Two individual panel datasets, identification (ID) panel and income panel, were derived from the IRP5 datasets for the period 2011–17. The ID panel links a unique ID number, representing a person, to each IRP5 certificate over the period. This allows the identification of individuals over multiple years. The income panel represents the aggregated income information for a specific individual (by unique ID number) for each tax year. This dataset therefore provides the total income of an individual, from multiple employers or from multiple IRP5 certificates from one employer, for each tax year.¹⁵

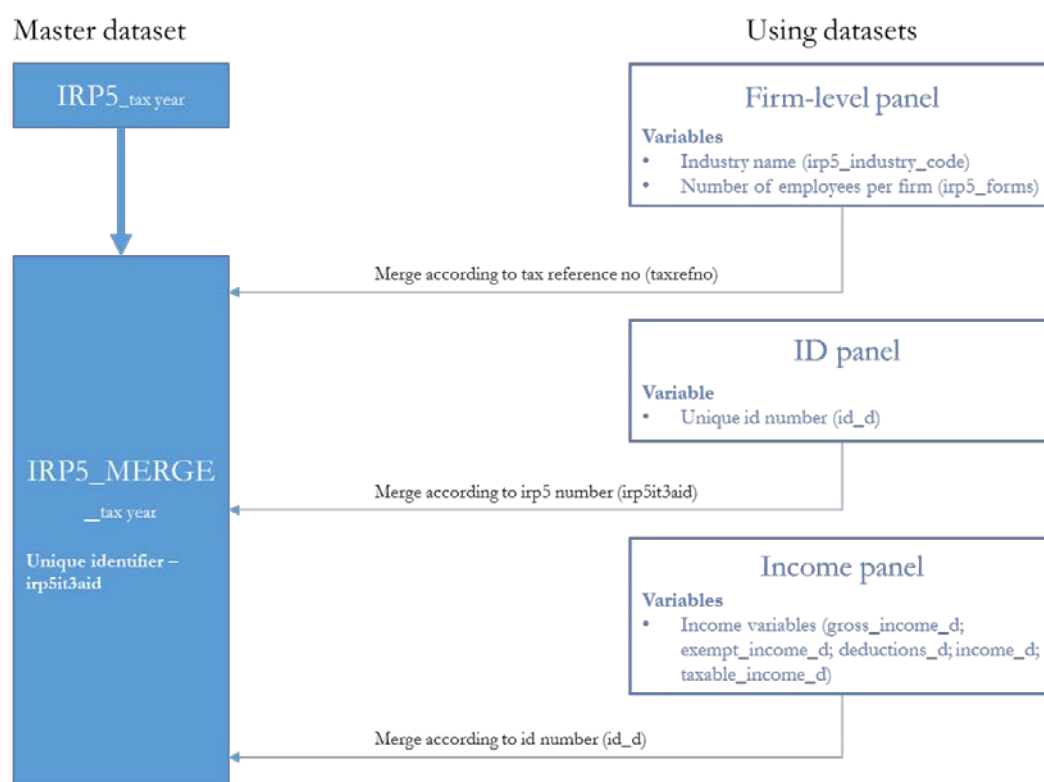
A firm-level panel dataset¹⁶ allows researchers to link the IRP5 data to specific company income tax (CIT) data such as the number of employees per firm and the firm industry.¹⁷ The individual panel datasets, as well as the firm-level panel dataset, were merged with the IRP5 datasets to construct a combined IRP5 dataset ('IRP5_Merge_201x') with each variable linked to a unique IRP5 number. This is illustrated in Figure 1.

¹⁵ Ebrahim and Axelson (2019) discussed the IRP5 and panel datasets in detail.

¹⁶ This was created by merging (1) company income tax data, (2) the IRP5 data (individual income tax), (3) value-added tax (VAT), and (4) customs records from traders. An introduction to these datasets was provided in Pieterse et al. (2018).

¹⁷ The VAT and customs data were not incorporated into the combined IRP5 dataset for this study.

Figure 1: Process of merging SARS datasets



Source: authors' construction.

Table 2 presents the main variables used in the research.

Table 2: Main variables in IRP5 datasets

Variable	Example	Description
<i>Identifiers</i>		
id_d	abcdek	Derived variable from ID panel. Identifies unique individuals over the tax years. Referred to as 'employees'.
irp5it3aid	1285012967	Unique IRP5 ID number allocated by SARS after a return is submitted. Indication of the number of jobs per employer or firm.
PAYE_RefNo	DJFKVMDL	The PAYE reference number of the IRP5 certificate. Referred to as 'employers'.
taxrefno	IEOLGMCS	Anonymized unique SARS tax reference number. Variable linked to the firm-level panel.
taxyear	2017	Refers to the tax year from 1 March to the last day of February (i.e. the 2017 tax year ran from 1 March 2016 to 28 February 2017).
<i>Demographics</i>		
Age	33	Derived from the date-of-birth variable. Age of the employee in the specific tax year.
Gender	M	Derived from the individual's South African ID number.
Nationality	1=ZA citizen, 0=Foreign	Derived variable indicating whether the employee is a South African citizen or not. Derived from whether an IRP5 observation is linked to an anonymized South African ID number or an anonymized passport number.
<i>Employment</i>		
employment_start	2016/03/01	Start date of employment in the particular tax year.
employment_end	2017/02/28	End date of employment in the particular tax year.

employment duration_days	365	Derived variable from employment_start and employment_end variables.
irp5_empl3601	5	Number of IRP5 forms receiving an income 3601 code* per firm in tax year.
id_d_Branch	4	Number of employees (id_d's) per employer (per PAYE reference number).
irp5_industry_code	MINING AND QUARRYING	Industry code derived from IRP5 form.
Gross_income_d	50293	From income panel. Sum of all income (business, normal, allowances, fringe benefits, lump sum income, etc.) for an individual in the specific tax year.
Nature_of_person	INDIVIDUAL	Only certificates on behalf of individuals were included. Other options represented in the IRP5 datasets included clubs, partnerships, retirement funds, and associations.
<i>Location</i>		
province_geo	Western Cape	Province in which the employee resides. Derived from postal address code.
districtmunicipal_geo	City of Cape Town	District municipality in which the employee resides. Derived from postal address code.
localmunicipal_geo	City of Cape Town	Local municipality in which the employee resides. Derived from postal address code.
mainplace_geo	Cape Metro	Main place in which the employee resides. Derived from postal address code.
busprov_geo	Western Cape	Province in which the employer (PAYE_RefNo) operates. Derived from postal address code.
busdistmuni_geo	City of Cape Town	District municipality in which the employer (PAYE_RefNo) operates. Derived from postal address code.
buslocmuni_geo	City of Cape Town	Local municipality in which the employer (PAYE_RefNo) operates. Derived from postal address code.
busmainplc_geo	Epping Industria	Main place in which the employer (PAYE_RefNo) operates. Derived from postal address code.

Note: * a 3601 code is an income source code that represents basic income on the IRP5 form.

Source: authors' construction.

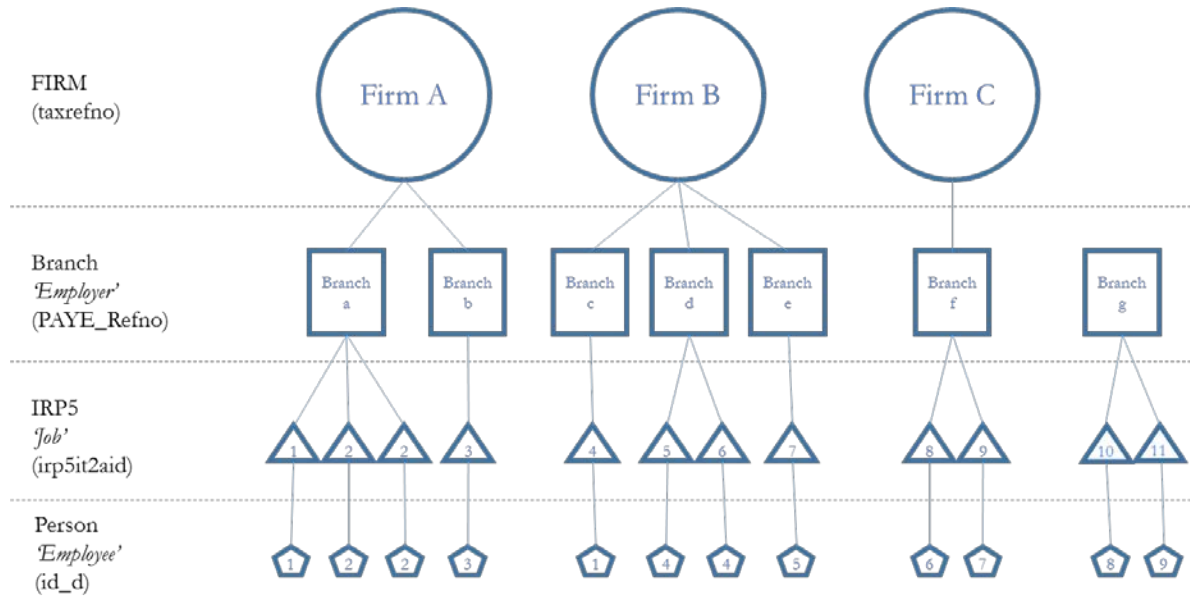
The unique identifier, 'id_d', in Table 2 is explained in detail in Pieterse et al. (2018: page 10). CIT information is available at the firm level and firms can have multiple branches. Take, for example, the Shoprite Checkers Supermarket Group, which is represented as a firm but consists of several different Shoprite and Checkers branches across the country. This study specifically examined employment at 'branch' level, which is represented by the unique PAYE reference number, and not at firm level, which represents companies in the CIT data. Branch-level data allowed travel distances to be derived between the employer at the branch level ('busmainplace_geo') and the employee's place of residence ('mainplace_geo').

Pieterse et al. (2018) also explained the linkage between firms, branches, jobs, and employees, and mentioned the different possible combinations in the SARS data. Figure 2 illustrates the different combinations and terminology used in this research. The following should be noted:

- There are branches and IRP5 forms that cannot be linked to firms (for example, Branch g), as some of the PAYE entities are linked to government institutions that do not pay CIT.
- In the IRP5 datasets, duplicate 'irp5it3aid' variables were observed, which were also linked to the same employee ('id_d'). These duplicate records were omitted in order not to overstate the number of jobs per branch (for example, 'Branch a' and 'IRP5 2').
- An employee ('id_d') can be associated with more than one branch: for example, where a consultant works for multiple companies simultaneously or where an individual has

worked for different employers for different periods of time in a specific tax year. This is represented as 'id_d 1' in Figure 2. An individual can also be linked to more than one IRP5 at the same company where an individual has worked for different time periods at the same company (see 'id_d 4').

Figure 2: Illustration of the relationship between firms, branches, IRP5 forms, and employees



Source: authors' construction.

The employment duration variable was derived from the 'employment_start' and 'employment_end' variables. These dates should not start before the first day of the tax year (1 March) and should also not end later than the last day of the end of the tax year (28 February, or 29 February in leap years). The procedure presented in Table 3 was followed for observations that did not meet these criteria to derive the variable employment 'duration_days'.

Table 3: Actions followed to derive employment duration variable

Variable	Recorded date	Action
'employment_start'	Within one month before the first day of the tax year (1 March)	Replace 'employment_start' record with 1 March (move the date forward to 1 March)
'employment_start'	More than one month before the first day of the tax year	Omit the observation
'employment_end'	Within one month after the last day of the tax year (28/29 February)	Replace 'employment_end' record with 28/29 February (move the date back to 28 or 29 February, depending on whether it is a leap year)
'employment_end'	More than one month after the last day of the tax year (28/29 February)	Omit the observation

Source: authors' construction.

Limitations of the location variables

Limitations to the methodology of creating spatial data are discussed in Ebrahim and Axelson (2019).

The spatial location of the main places where the employee resides (home) and where the employer is located (workplace) was derived by obtaining the centroid of the postal code reflected on the IRP5 form. By running the postal codes through the Google application programming interface (API), the matching main place was identified (Ebrahim and Axelson 2019). This centroid was then linked to a specific main place in which that centroid fell, and the main place was linked to the more aggregated location fields such as the local and district municipality and province.

The limitations to this approach are as follows:

- The Google API could have allocated an incorrect output of the postal code centroid.
- The centroid, which is allocated by the Google API, may not accurately reflect the centroid of the postal code, which is easily distorted when working with non-circular polygons.
- The postal code zone may overlap several main places, and the centroid of the postal code may not fall within the main place which covers the largest area of the postal code.
- The address field of the employee is very poorly populated, as it is not compulsory to complete this field on the IRP5 form. The workplace addresses have been populated more consistently since 2013. Table 4 indicates the percentage of residential and workplace location information obtained for the specific tax year. In addition, it indicates the percentage of observations that have both origin and destination data available.

These limitations could lead to incorrect distance calculations due to attributing the wrong main place centroid to an observation. However, these biases will occur less in metropolitan areas, where postal codes are better geographically defined and main places smaller compared with rural areas.

Table 4: Percentage of observations in the IRP5 datasets with location data

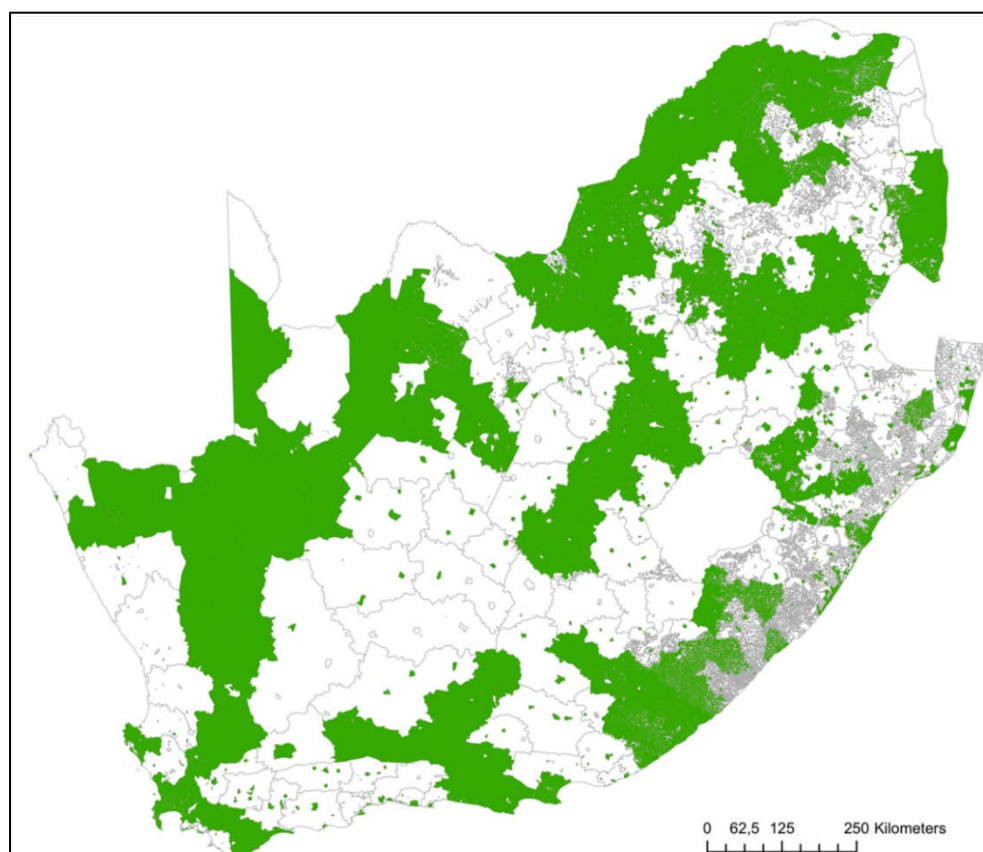
Tax year	Residential	Workplace	O-D pairs
2011	21	1.30	0.3
2012	34	5.20	1.8
2013	42	73	42
2014	32	90	31
2015	16	91	15
2016	16	90	16
2017	16	88	16

Source: authors' construction based on from Ebrahim and Axelson (2019) and IRP5_Merge_201x datasets

The SARS IRP5 datasets from 2013 to 2017 were used for further analysis in this research as they contained the highest sample of observations with O-D pairs as indicated in Table 4.

Figure 3 illustrates in green the main places where data exist on where workers reside and where they work. Large parts of rural South Africa are not represented in the data, but there is good representation of metropolitan areas. This figure also indicates the size of the main places, which is the most disaggregated spatial unit for the location data in the SARS IRP5 datasets. The coarseness of this zoning system may lead to inter-zonal travel patterns not being observed or could lead to problems such as the modifiable areal unit problem (MAUP) whereby data are totalled to a more aggregated spatial unit, which may influence the results of statistical hypothesis testing (Openshaw 1983).

Figure 3: Location data available by main place (2017 tax year)



Source: authors' construction based on IRP5_Merge_2017 dataset.

Representativeness of the sample

Table 4 indicates the percentage of observations with O-D pairs that may result in the sample of observations with O-D pairs being unrepresentative of the full sample. Table 5 indicates the percentage and mean difference in the full sample compared with the O-D sample for different variables across years. It can be seen that the variables 'age', 'gender', and 'nationality' are not significantly different between the full sample and the O-D sample. The mean gross income variable, however, indicates that the O-D sample is on average 10 per cent higher than the full sample. This indicates that the O-D sample under-represents lower-income groups. This under-representation should be kept in mind when interpreting the results.

Table 5: Difference in sample statistics for the variables 'age', 'income', 'gender', and 'nationality'

Variables	2017		2016		2015		2014		2013	
	All	O-Ds	All	O-Ds	All	O-Ds	All	O-Ds	All	O-Ds
Total observations ('000)	13,759	2,200	13,574	2,128	13,655	2,091	13,305	4,124	11,319	4,755
% observations kept (with O-Ds)		16		16		15		31		42
Age (mean)	37.4	36.8	37.3	36.7	37.2	36.8	37.1	38.2	36.5	37.1
Age (mean) difference (in years)		-0.5		-0.6		-0.4		1.1		0.6
Gross income (mean) ('000) (ZAR)	204	231	179	209	196	196	179	198	135	161
Gross income (mean) (%) difference		11		14		0		10		16
Male (%)	0.53	0.56	0.53	0.56	0.54	0.57	0.54	0.51	0.57	0.56
Male (proportion difference)		0.03		0.03		0.04		-0.03		-0.01
Foreign (%)	0.03	0.04	0.03	0.04	0.03	0.03	0.03	0.02	0.03	0.03
Foreign (proportion difference)		0.00		0.01		0.01		-0.01		0.00

Source: authors' construction based on IRP5_Merge_2017 dataset.

A process of data cleaning and transforming was followed to derive a cleaned O-D dataset for the 2013–7 tax years. The following tasks were undertaken on the datasets:

- Where duplicate 'irp5it3aid' existed in the same branch ('PAYE_Refno') linked to the same employee ('id_d') in a tax year, all duplicates were omitted, with only one entry remaining.
- All records were omitted if the IRP5 record indicated that the nature of the person ('Nature_of_person' variable) was other than 'INDIVIDUAL'. This was done to exclude trusts, partnerships, etc. from the datasets.
- If an individual was stated to be younger than 15 in the particular tax year (15 is the youngest age at which a person may legally be employed in South Africa), the observation was omitted. An observation was also omitted if a person was older than 75.
- If the employment duration field of the observation was missing, the observation was omitted.
- If an employee was linked to more than one 'irp5it3aid' record in the datasets for the same employer ('PAYE_Refno'), only the record with the longest employment duration for that individual at the specific employer was retained and the other duplicate observations were omitted.
- Observations with a gross annual income of zero or less (2.5 per cent of observations) were omitted.

A geographical classification was generated for each observation to indicate whether the employer (workplace) is situated in a metropolitan area, urban area, or rural area. These classifications were derived from the 2011 Census data classifications by EA. The following procedure was followed to generate this variable:

- If the municipal type ('mn_type' variable in Census 2011 data) was 'Metro', the main place that fell within that municipality was classified as 'Metro'.
- If an EA was classified as 'Farms' or 'Traditional' (the 'EA_gtype' variable in the Census 2011 data), the EA was classified as 'Rural'.
- If an EA was classified as 'Urban' (the 'EA_type' variable in the Census 2011 data), the EA was classified as 'Urban'.

A main place consists of many EAs, resulting in main places with different EA geographical classifications. The total square kilometres for each classification according to 'Metro', 'Urban', or 'Rural' was calculated by main place, and the classification that covered the highest percentage area within the main place was selected.

Spatial mapping was conducted to investigate the employment and travel characteristics at metropolitan level using ArcMAP. Descriptive statistics in the form of histogram plots and tables were created to provide insights into the employment and travel characteristics of different income groups, individuals commuting in different metropolitan areas, and different geographical classifications.

Correlation analysis, pairwise mean comparisons, and regression analysis were conducted to investigate the relationships between travel distance, transport proximity, and employment duration. These analyses were conducted by excluding all trips of over 100 km to prevent long distances such as trips between metropolitan areas from distorting the outcomes. The deciles for airline distances from 2013 to 2017 were calculated to determine the most appropriate cut-off point. The summarized statistics (see Table 6) indicate that at least 80 per cent of trips were shorter than 100 km over the five-year period, with average distances varying between 110 km and 114 km. This resulted in the decision to choose a 100 km cut-off over all five years.

Table 6: Airline distance deciles and mean airline distance between 2013 and 2017 tax years

Airline distance deciles	2017	2016	2015	2014	2013
10	3.037316	2.83303	2.846097	3.469148	2.983863
20	5.855942	5.854458	5.854458	6.863139	5.415972
30	8.242757	7.987182	8.38528	8.640365	8.13919
40	10.02028	10.02028	10.19311	12.19849	10.56262
50	13.33793	13.2504	13.33793	17.48068	16.15635
60	20.46617	19.63257	19.47487	27.99332	26.459
70	35.26675	33.96829	34.03584	50.7902	47.91126
80	74.72463	68.74125	68.74125	110.454	101.7226
90	372.1665	344.3791	344.7335	339.2555	336.4383
Mean distance	114	110	111	114	111

Source: authors' construction based on IRP5_Merge_2017 dataset.

5 Data analysis

5.1 Job–housing balance

Figures 4 to 6 indicate the mismatch between where workers in South Africa reside and where they are employed within the five main metropolitan areas¹⁸ according to the IRP5_Merge_2017 dataset. They also indicate the working-class population¹⁹ according to the 2011 Census by main place. This mismatch is represented by a worker/employment ratio (see Table 7).²⁰ A ratio smaller than 1 indicates higher numbers of employment opportunities than the number of workers residing in the main place, resulting in an inflow of workers to that main place. A ratio larger than 1 indicates higher numbers of workers residing in the main place than the number of employment opportunities, which indicates an outflow of workers from the main place. These figures also indicate the average airline distance²¹ for workers travelling to a specific main place.

Table 7: Worker/employment opportunity ratio by metropolitan

Metropolitan	Worker/employment opportunity ratio				Airline distance
	<i>Mean</i>	<i>Std. dev</i>	<i>Min.</i>	<i>Max.</i>	<i>Mean (km)</i>
City of Cape Town	4.01	7.330359	0.15	31	15
Gauteng region	6.2	6.203325	0.32	35	22
eThekweni	9.01	28.41567	0.28	188	10

Source: authors' construction based on IRP5_Merge_2017 dataset.

The City of Cape Town shows a high inflow of workers to the main central business district (CBD), City Bowl, with a relatively low number of workers who reside in the same main place. This indicates that most workers must travel between 15 and 20 km to reach employment opportunities in this main place. Figure 4 also indicates a surplus of workers who reside in the Metro South-East (Philippi, Mitchells Plain, and Khayelitsha), with limited work opportunities in these predominately low-income communities. Similar results were reported by Krygsman et al. (2016), which indicated large employment surpluses in the city centre and residential surpluses in predominately low-income residential areas.

eThekweni shows a better job–housing balance compared with the other two metropolitan areas. Figure 5 indicates that most workers are employed in the Durban main place, which requires commuters to travel between 15 and 20 km to reach their place of employment. The Gauteng metropolitan area shows a large inflow of workers into the Johannesburg, Sandton, and Pretoria CBDs, with a noticeably large outflow of workers from the Soweto main place. Figure 6 indicates that people travel the longest distances in this metropolitan area compared with Cape Town and eThekweni, where most workers travel between 20 and 30 km to their place of employment.

Travel time is a function of network distance, speed, and interconnectivity. As Kgwedi and Krygsman (2019) indicated, travel using public transport, specifically rail and bus, involves

¹⁸ The City of Cape Town, eThekweni, and Gauteng region, the latter representing the City of Johannesburg, the City of Tshwane, and Ekurhuleni combined.

¹⁹ All individuals between the ages of 15 and 65.

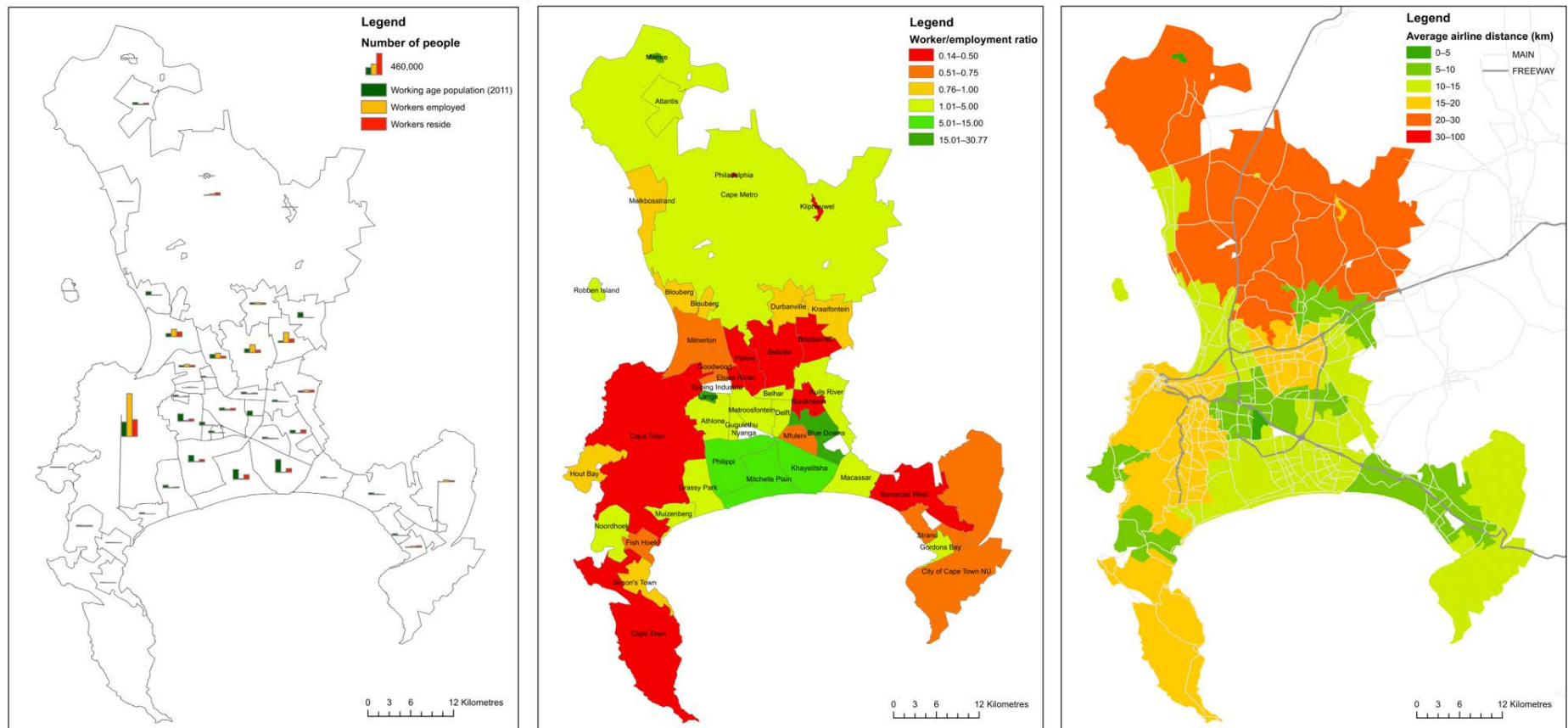
²⁰ Calculated by dividing the total number of workers who reside in that main place by the total number of workers employed in that main place.

²¹ It is important to note that Figures 4 to 6 (right-hand graphs) represent airline distance and not actual travel time or distance.

significant transfers and additional access and egress stages. Low-income communities may face significantly longer travel times, as revealed by airline network distance.

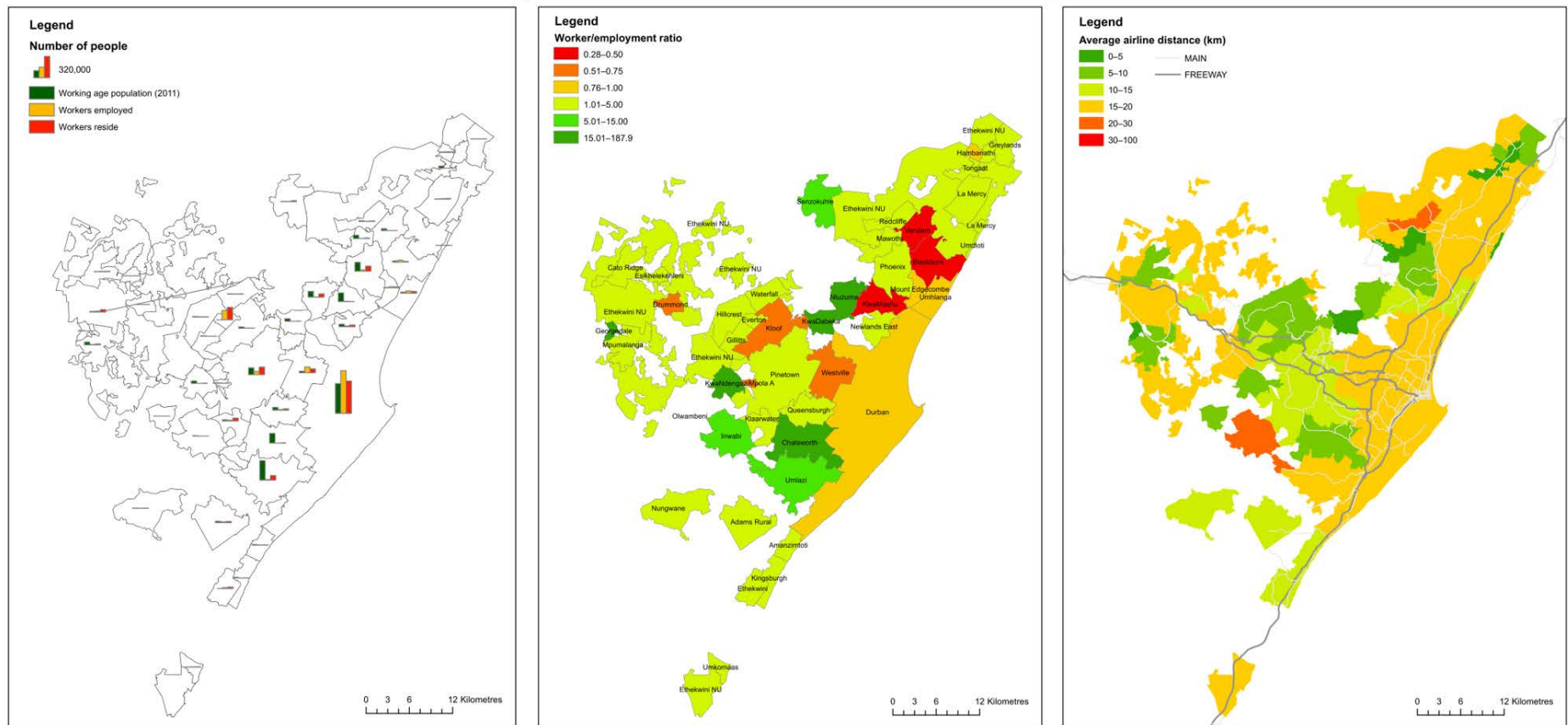
These figures indicate the extent of the job–housing imbalance in South African cities and the need to stimulate small and medium enterprises and other work opportunities in areas where people reside in order to reduce the distances travelled by workers. Attention should also be paid to increasing investment in public transport from the main places where a surplus of workers reside to the main places where there is a surplus of work opportunities.

Figure 4: City of Cape Town: Population and employment distribution and average airline distance to employment (2017 tax year)



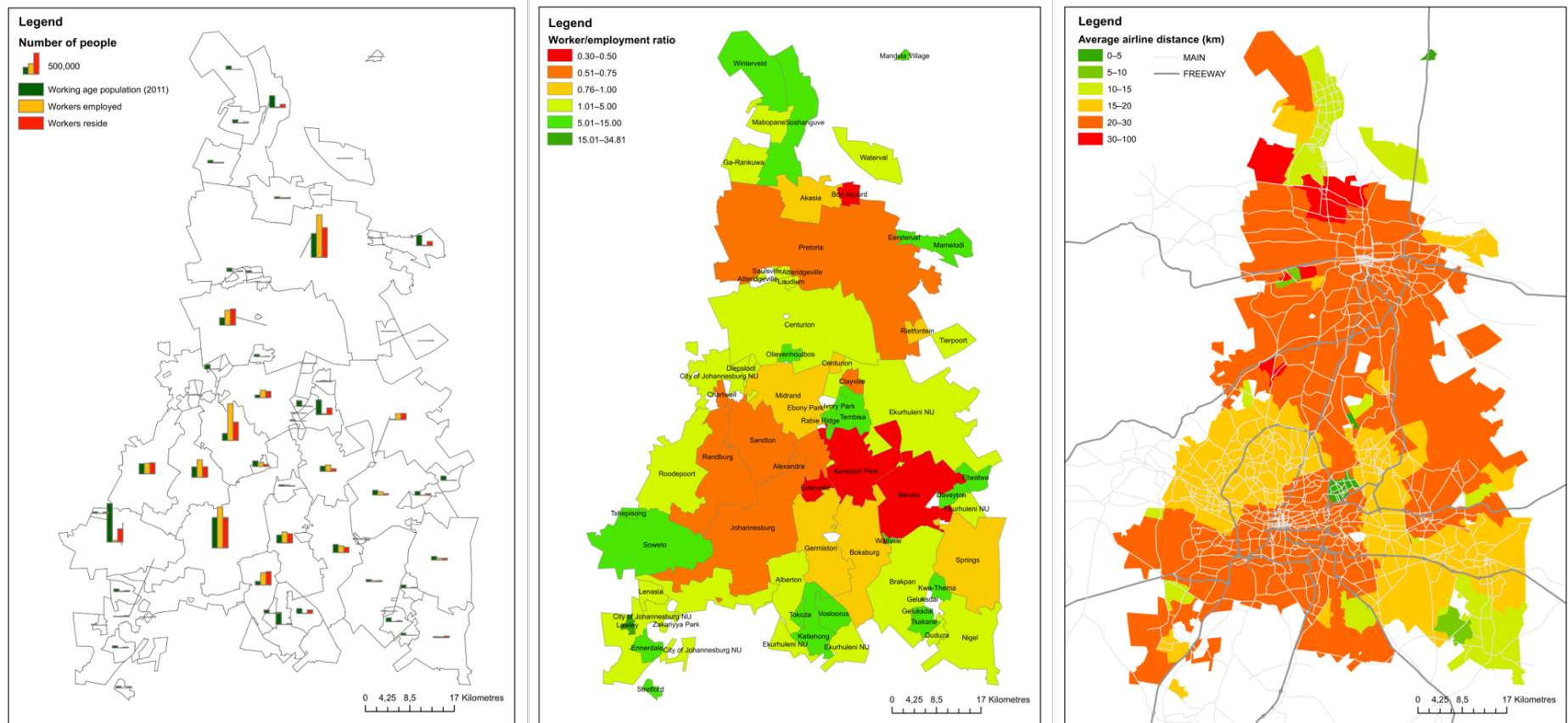
Source: authors' construction based on aggregated main place level (2017 tax year) data ('density_ZA_2017.dta').

Figure 5: eThekweni: Population and employment distribution and average airline distance to employment (2017 tax year)



Source: authors' construction based on aggregated main place level (2017 tax year) data ('density_ZA_2017.dta').

Figure 6: Gauteng: Population and employment distribution and average airline distance to employment (2017 tax year)



Source: authors' construction based on aggregated main place level (2017 tax year) data ('density_ZA_2017.dta').

5.2 Descriptive analysis of employment duration and airline distance

Employment duration

Figure 7 shows the cumulative employment duration. The graph indicates a gradual increase in employment duration for employees ranging between 1 day and 11 months. After the 11th month of employment, there is a steep increase in the percentage of individuals who work for between 11 and 12 months within the specific tax year. In 2017, more than 60 per cent of individuals remained employed for longer than nine months of the year.

Figure 7: Cumulative employment duration of employees in the 2017 tax year



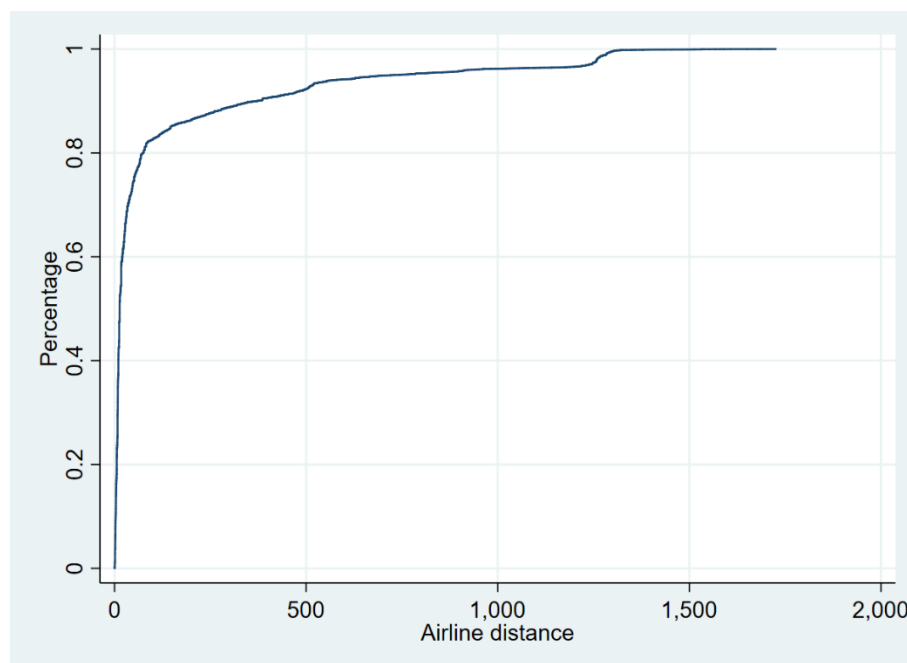
Note: data also available for tax years 2013–16.

Source: authors' construction based on IRP5_Merge_2017 dataset.

Airline distance

Figure 8 provides an indication of individuals' proximity to their place of employment (measured in airline distance). The inflection points on the cumulative graph indicate that approximately 75 per cent of all employees live closer than 50 km away from their workplace. The slope declines after 50 km, with two small inflection points at approximately 500 km and 1,300 km. These inflection points most probably reflect the trips made between Durban and Johannesburg, an approximate 500 km airline distance apart, and between Cape Town and Johannesburg, an approximate 1,300 km airline distance apart. It is not possible to identify from the SARS data whether the individual works in person at the physical place of employment. It is assumed that many of the records that indicate individuals who travel long distances are attributed to consultancy work that allows for remote work and does not require the employee to always be physically present at the place of employment.

Figure 8: Cumulative airline distance between an employee's residence and workplace (2017 tax year)



Note: data also available for tax years 2013–16.

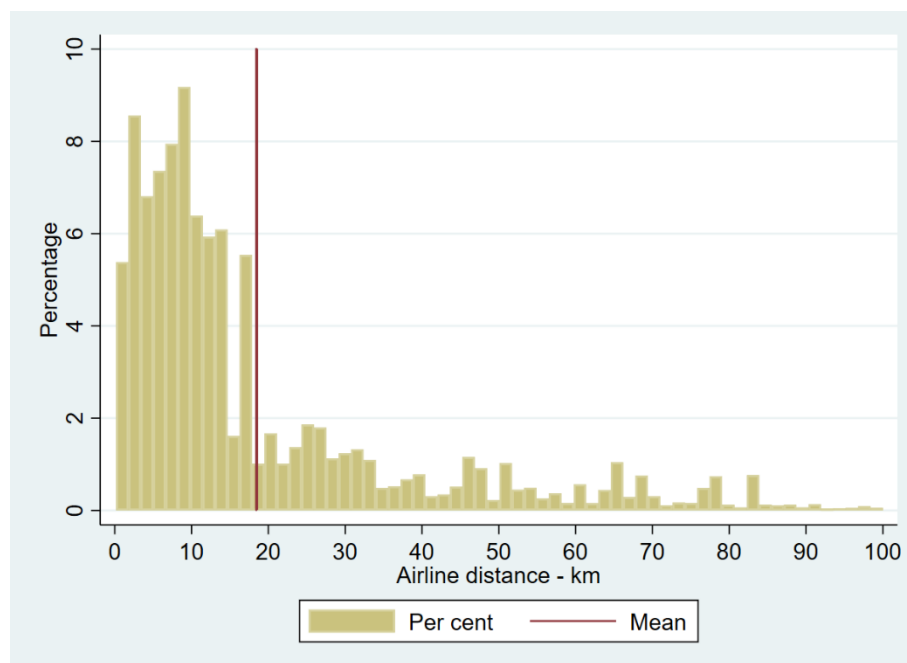
Source: authors' construction based on IRP5_Merge_2017 dataset.

Considering the short-distance trips and O-D pairs with an airline distance of 100 km and less reveals that most trips are below 20 km, with a mean of 18 km, as indicated in Figure 9.

In order to explore the influence of distance on employment duration, a pairwise mean comparison was conducted between airline distances for workers in different income groups. Table 8 indicates that the mean airline distance of lower-income commuters is significantly²² shorter than that of higher-income commuters. Higher-income households, with access to private transport, have a high mean airline distance of 22 km. Because they rely on private transport, they can access significantly more employment opportunities. On the other hand, the employment search space for low-income categories is significantly smaller, i.e. 14 to 16 km. The shorter mean airline distances for low-income communities may in fact reflect their poor access to employment. This is confirmed by the outputs of the NHTS 2013 data, indicating that between 80 and 90 per cent of trips made by low-income commuters are made by public transport or by walking all the way, compared with only 40 per cent of high-income commuters (Stats SA 2015).

²² At a 5 per cent significance level.

Figure 9: Frequency of trips by airline distance between an employee's residence and workplace (2017 tax year)



Note: data also available for tax years 2013–16.

Source: authors' construction based on IRP5_Merge_2017 dataset.

Table 8: Mean airline distance by income decile (2017 tax year)

Income decile	Income range (ZAR)	Airline distance (km) mean	Standard error	Bonferroni groups*
Income decile 2	1–10,164	14.25425	0.043833	A
Income decile 1	10,165–25,799	14.42094	0.043483	A
Income decile 3	25 800–41,458	15.59084	0.043626	
Income decile 4	41,459–59,537	16.08015	0.044278	
Income decile 5	59,538–87,099	17.31551	0.04513	
Income decile 6	87,100–137,942	18.81981	0.046095	
Income decile 7	137,943–208,283	22.39262	0.047816	B
Income decile 8	208,284–316,926	22.52404	0.047336	B
Income decile 10	316,928–533,068	22.7848	0.046868	C
Income decile 9	533,073–121,000,000	22.79335	0.047817	C

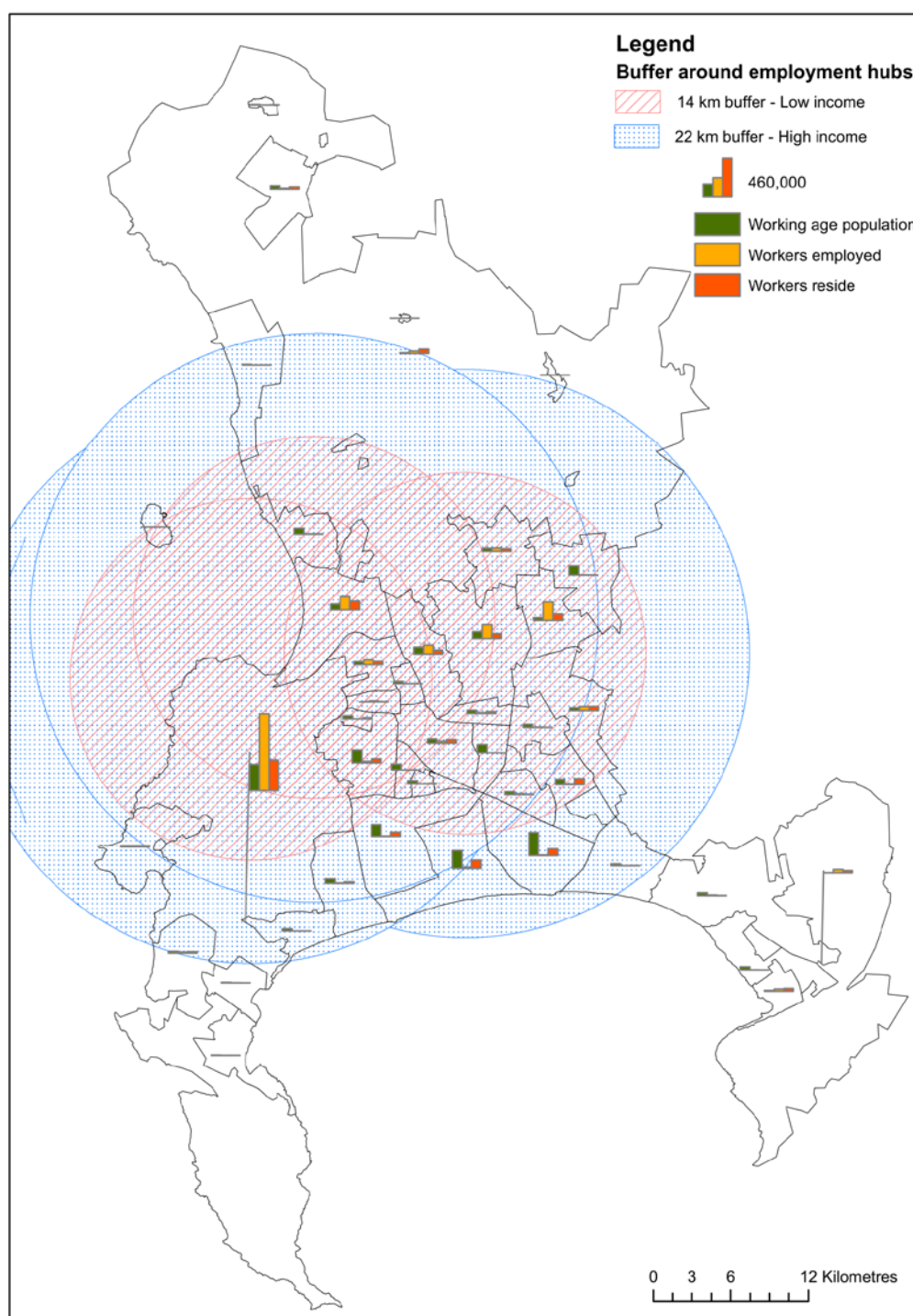
Note: * means sharing a letter in the group label are not significantly different at the 5% level.

Source: authors' construction based on IRP5_Merge_2017 dataset.

As we will show below, the absolute number and choices of employment opportunities are much lower for these income categories. The short airline distance for low-income residents is therefore not an indication of a good location for employment, but rather a reflection of their limited ability to access employment using public transport. Whether this is also reflected in employment duration is discussed in Section 5.3.

The employment search space for the City of Cape Town is illustrated in Figure 10, which indicates the limited access to employment opportunities of lower-income individuals who reside in the Metro South-East.

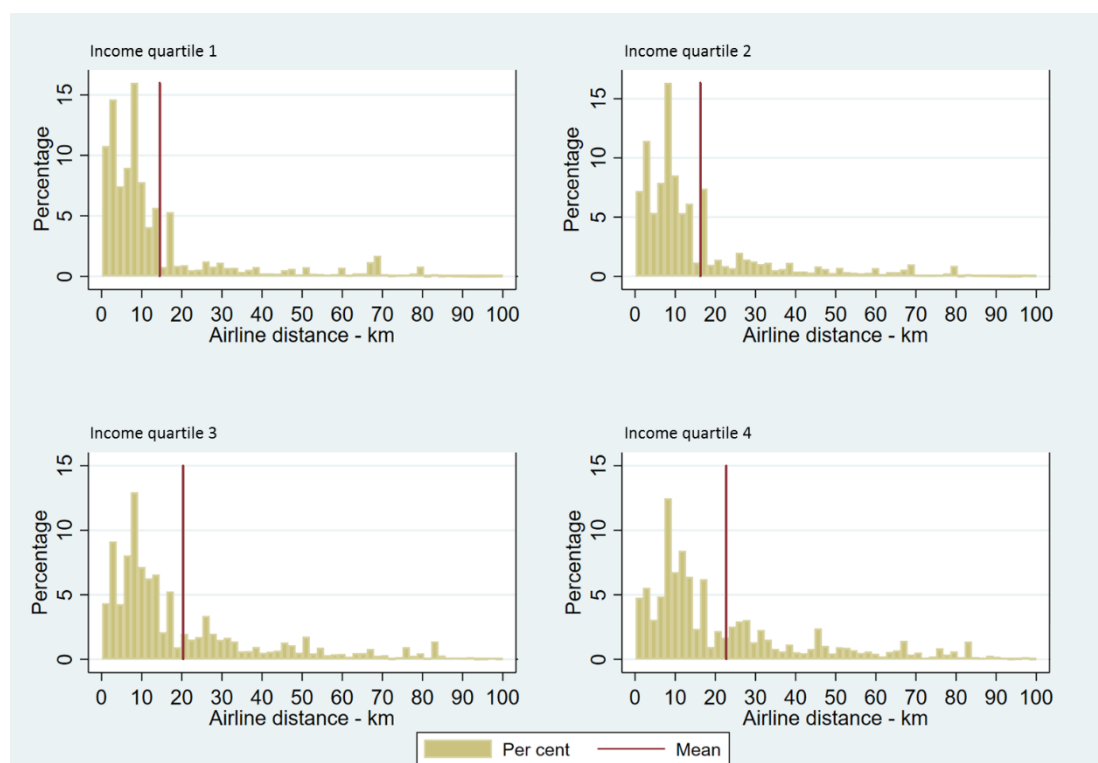
Figure 10: Employment search space for the City of Cape Town (2017 tax year)



Source: authors' construction based on aggregated main place level (2017 tax year) data ('density_ZA_2017.dta').

Figure 11 indicates the different airline distances by income quartiles. It shows that lower-income groups travel shorter distances (average of 15 km), with very few low-income workers residing more than 30 km from their workplace. In comparison, higher-income workers (income quartiles 3 and 4) travel 22 km on average.

Figure 11: Frequency by airline distance between an employee's residence and workplace for different income quartiles (2017 tax year, trips shorter than 100 km)



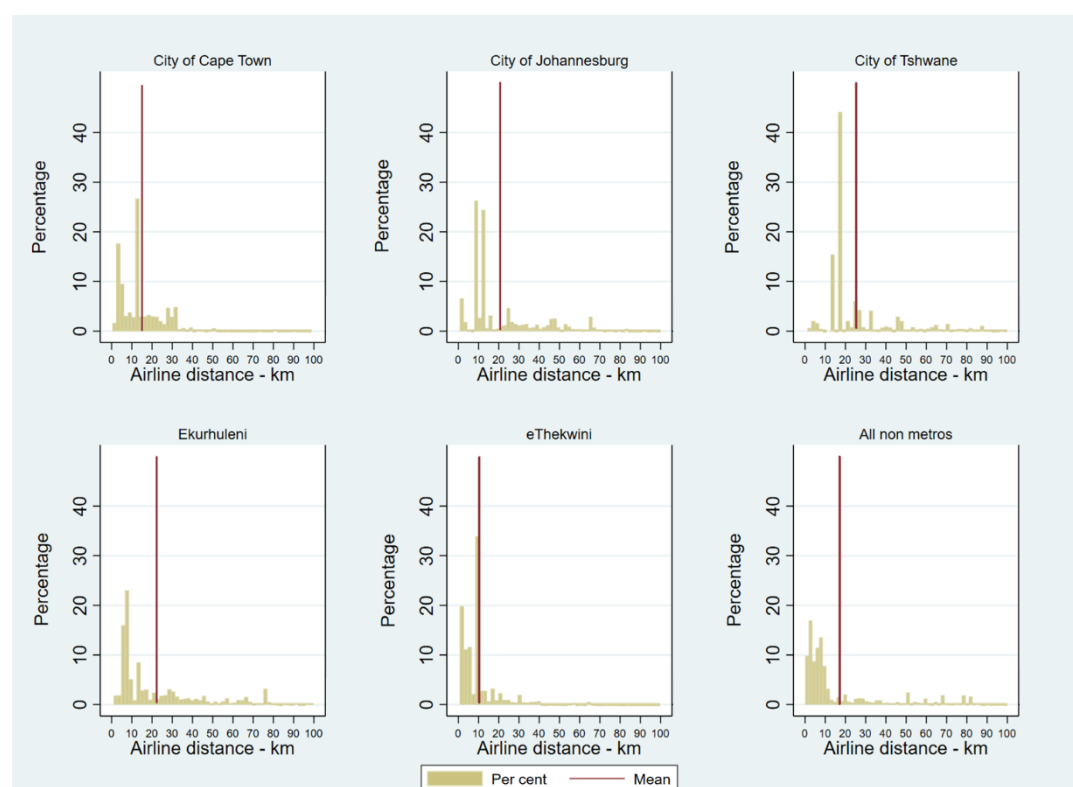
Notes: income Q1 = gross_income \leq ZAR33,953 per year; income Q2 = gross_income $>$ ZAR33,953 and gross_income \leq ZAR87,099 per year; income Q3 = gross_income $>$ ZAR87,099 and gross_income \leq ZAR254,182 per year; income Q4 = gross_income $>$ ZAR254,182. Data also available for tax years 2013–16.

Source: authors' construction based on IRP5_Merge_2017 dataset.

Figure 12 indicates the airline distance profiles of employees who travel to their place of employment for the major metropolitan²³ areas in South Africa. The figure indicates that the average airline distances for commuters in Cape Town and eThekwin are lower, with mean airline distances of 15 km and 10 km respectively. Commuters who work in the City of Tshwane travel on average the longest, with an average airline distance of 25 km.

²³ The metropolitan represents the metro in which the employer is situated, not necessarily where the employee resides.

Figure 12: Frequency by airline distance between an employee's residence and workplace for different metropolitans (2017 tax year, trips shorter than 100 km)

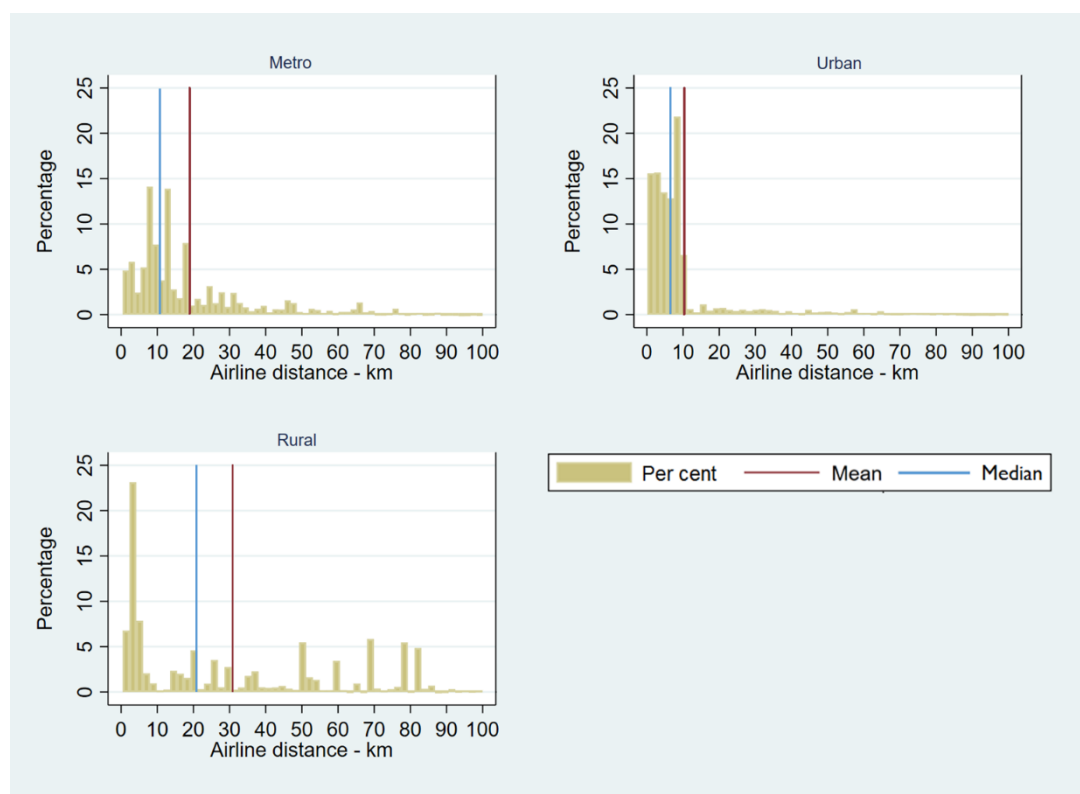


Note: data also available for tax years 2013–16.

Source: authors' construction based on IRP5_Merge_2017 dataset.

Figure 13 indicates the airline distance according to the different geographical classifications and shows significant differences between metropolitan, urban, and rural areas. It indicates that people in rural areas either commute less than 5 km (which makes non-motorized transport a viable mode of choice) or must travel very great distances relative to metropolitan or urban areas. The figure also indicates that 50 per cent of individuals travel more than 20 km in rural areas compared with the median of 10 km for individuals employed in urban or metro areas.

Figure 13: Frequency by airline distance between an employee's residence and workplace for different geographical classifications (2017 tax year, trips shorter than 100 km)



Note: data also available for tax years 2013–16.

Source: authors' construction based on IRP5_Merge_2017 dataset.

5.3 Correlation analysis, pairwise mean comparisons, and regression analysis

In this section we extend the descriptive analysis by investigating the relationship between airline distance and employment duration controlling for the influence of other relevant variables.

Table 9 shows that the linear relationship between airline distance and employment duration (measured in days) is weak but significant (at a 5 per cent significance level). An increase of 1 km in distance will result in an increase of 0.0685 days' employment duration for all income groups. This finding appears counterintuitive and rejects the hypothesis that increased travel distance will result in lower employment duration. But, as we showed above, there is a complex relationship between income quartiles and employment, and here too the coefficients change signs when differentiating between income quartiles. This indicates that there is a significant negative linear relationship between employment duration and airline distance for income quartiles 1 and 2 (lower-income groups) and a positive relationship for income quartiles 3 and 4 (higher-income groups).

As discussed, this may reflect the employment search space dynamics. The more the employment search space increases, the more likely individuals will be to access higher-paying jobs and jobs they prefer (in terms of location, flexibility, other non-monetary benefits, etc.). The flexibility of the private vehicle will facilitate this larger employment search space. Individuals will therefore be more likely to stay employed longer in these jobs, which supports the positive relationship between airline distance and employment duration. The negative coefficients for low-income categories contrast with the higher-income results. As distance increases, transport costs increase (travel time and price of the trips). As the 'costs of employment' increase with commuting costs, lower-income households may not be able to absorb these costs, which results in shorter employment duration.

These relationships remain weak, but do, however, indicate that travel distance has a negative impact on employment duration for lower-income groups. Other factors may influence employment duration for high-income groups.

Table 9: Pairwise correlation between employment duration (days) and airline distance by income quartiles (2017 tax year)

Pairwise correlation	All income groups	Income Q1	Income Q2	Income Q3	Income Q4
Airline distance (km)	0.0685	-0.0270	-0.0118	0.0394	0.0138

Notes: income Q1 = gross_income ≤ ZAR33,953 per year; income Q2 = gross_income > ZAR33,953 and gross_income ≤ ZAR87,099 per year; income Q3 = gross_income > ZAR87,099 and gross_income ≤ ZAR254,182 per year; income Q4 = gross_income > ZAR254,182. Only includes trips with an airline distance of 100 km or less. Significant at 5% level.

Source: authors' construction based on IRP5_Merge_2017 dataset.

Table 10 compares employment duration for employees who live and work in the same main place with duration for those who travel between main places, by different income quartiles. It is assumed that individuals who live and work in the same main place will have better accessibility than individuals who travel between main places. The results indicate a 16-day difference in mean employment duration for individuals who fall into the lowest income quartile and an eight-day difference for commuters in the second income quartile. For income quartiles 3 and 4, however, the results reinforce the findings of Table 9, and the mean employment duration increases when travel distance increases. This may be a result of higher-income earners having access to private vehicle usage, which allows them to access better job opportunities even though these may be at a greater distance. The results again indicate that higher- and lower-income employees have different characteristics in terms of employment duration and distance. While higher-income individuals—predominately private car users—can access more, and a wider choice of, employment opportunities, they are more likely to travel further to their workplace.

The result confirms the hypothesis that increased transport accessibility, measured here by whether an individual resides and works in the same main place, results in higher employment duration for lower-income groups.

Table 10: Pairwise mean test between employment duration and whether an employee lives and works in the same main place by income quartile (2017 tax year)

Employment duration (days)		Mean (employment duration)			
		Income Q1	Income Q2	Income Q3	Income Q4
Lives and works in different main place	MP_OD_equal==0	150,977	292,179	313,269	321,325
Lives and works in same main place	MP_OD_equal==1	166,980	300,377	303,025	316,053
P > t	MP_OD_equal: 1 vs 0	0.000*	0.000*	0.000*	0.000*

Note: * significant at 5% level.

Source: authors' construction based on IRP5_Merge_2017 dataset.

A regression analysis was conducted to analyse the relationship between individual demographic characteristics, employment, transport accessibility, and location variables in the IRP5_Merge_2017 dataset on employment duration. Table 11 describes the variables used within this regression analysis.

Table 11: Regression analysis variable description

Variable	Description
MP_OD_equal	Dummy variable: 1 = individual lives and works in same main place; 0 = individual lives and works in a different main place
air_dist_km_Internal	Continuous variable: airline distance in kilometres
gross_income_d	Continuous variable: annual gross income
age_d	Continuous variable: age
male	Dummy variable: 1 = male; 0 = female
foreign	Dummy variable: 1 = foreigner; 0 = South African
id_d_Branch	Number of workers employed by employer (indication of branch size)
employed_2016	Dummy variable: 1 = individual was employed by the same firm in 2016; 0 = individual was not employed by the same firm in 2016
employed_2015	Dummy variable: 1 = individual was employed by the same firm in 2015; 0 = individual was not employed by the same firm in 2015
employed_2014	Dummy variable: 1 = individual was employed by the same firm in 2014; 0 = individual was not employed by the same firm in 2014
geo-type_Urban_dum_Bus	Dummy variable: 1 = individual works in a main place classified as 'urban'. Reference category: 'Metro'
geo-type_Rural_dum_Bus	Dummy variable: 1 = individual works in a main place classified as 'rural'. Reference category: 'Metro'
_cons	constant

Source: authors' construction.

The analysis was conducted for all income groups (see Table 12) and for the different income quartiles (see Table 13). In Tables 12 and 13, all coefficients are significant at a 5 per cent confidence level.

The first regression analysis (over all income groups) model explains 31 per cent of the variation in the data. Other variables, which cannot be measured using the tax administrative data (such as education level), may have explained a higher percentage of the variation in the independent variable. The model shows that employment duration will have a positive relationship with the individual living and working in the same main place. However, analysing the results over different income groups (see Table 12) indicates that for the lowest income quartile, individuals work on average 23 days longer when they live and work in the same main place.

Both tables indicate a negative relationship between airline distance and employment duration: an increase of 1 km in commuting distance will decrease employment duration by between 0.05 and 0.24 days, depending on income group. The tables also show that lower-income groups are more affected by airline distance than higher-income groups.

Overall, employment duration increases relative to an increase in the gross income of individuals, as well as an increase in age. In addition, individuals who were employed by the same employer in the previous tax year will have a higher employment duration than those who were not.

Lastly, the geo-classification variables show that for all income groups, the employment duration decreases if an individual works in an urban or rural area compared with working in a metropolitan area. However, the sign of the relationship changes for different income quartiles, with higher-income earners increasing their employment duration when working in urban and rural areas compared with metropolitan areas.

Table 12: Regression analysis, all income groups (2017 tax year)

Source	SS	df	MS	Number of obs	=	1,814,793
Model	8.4557e+09	13	650,435,567	F(13, 1814779)	=	61,375.63
Residual	1.9232e+10	1,814,779	10,597.6186	Prob > F	=	0
Total	2.7688e+10	1,814,792	15,256.8438	R-squared	=	0.3054
				Adj R-squared	=	0.3054
				Root MSE	=	102.94

<i>Employment duration (days)</i>	<i>Coef.</i>	<i>Std. err.</i>	<i>t</i>	<i>P > t </i>	<i>[95% conf. interval]</i>	
MP_OD_equal	-10.83829	0.198256	-54.67	0.000	-11.22686	-10.44971
air_dist_km_Internal	-0.1238486	0.0047933	-25.84	0.000	-0.1332434	-0.1144539
gross_income_d	6.70e-06	1.52e-07	44.11	0.000	6.40e-06	7.00e-06
age_d	0.6715147	0.0075841	88.54	0.000	0.6566502	0.6863792
male	-1.055464	0.1579183	-6.68	0.000	-1.364979	-0.74595
foreign	15.78631	0.4136219	38.17	0.000	14.97562	16.59699
id_d_Branch	0.0002071	3.25e-06	63.69	0.000	0.0002008	0.0002135
employed_2016	100.3308	0.2161647	464.14	0.000	99.90716	100.7545
employed_2015	28.61857	0.2679534	106.80	0.000	28.09339	29.14375
employed_2014	10.26319	0.3231524	31.76	0.000	9.629818	10.89655
employed_2013	14.15443	0.2857633	49.53	0.000	13.59434	14.71452
geo_type_Urban_dum_Bus	-11.10036	0.194553	-57.06	0.000	-11.48168	-10.71904
geo_type_Rural_dum_Bus	-0.5895572	0.2610431	-2.26	0.024	-1.101193	-0.0779218
_cons	159.3533	0.3464686	459.94	0.000	158.6743	160.0324

Source: authors construction based on IRP5_Merge_2017 dataset.

Table 13: Regression analysis, different income quartiles (2017 tax year)

Employment duration (days)	Income Q1		Income Q2		Income Q3		Income Q4	
	<i>Coef.</i>	<i>P > t </i>	<i>Coef.</i>	<i>P > t </i>	<i>Coef.</i>	<i>P > t </i>	<i>Coef.</i>	<i>P > t </i>
MP_OD_equal	23.25908	0.000	0.9755803	0.010	-6.984823	0.000	-1.992595	0.000
Airline distance (km)	-0.244461	0.000	-0.0523061	0.000	-0.0569209	0.000	-0.0681758	0.000
Gross annual income	0.00487	0.000	-0.0000769	0.000	0.0000845	0.000	-3.14e-06	0.000
age_d	0.3755614	0.000	0.1894031	0.000	-0.0730439	0.000	-0.1942397	0.000
male	-22.10546	0.000	-5.308701	0.000	-1.243935	0.000	-0.5209832	0.058
foreign	4.929321	0.000	17.20274	0.000	16.04656	0.000	20.70094	0.000
id_d_Branch	0.001317	0.000	-0.0001908	0.000	0.0000934	0.000	0.0001448	0.000
employed_2016	42.34035	0.000	99.04895	0.000	95.91547	0.000	85.40441	0.000
employed_2015	5.255627	0.000	11.65029	0.000	17.62068	0.000	31.99551	0.000
employed_2014	12.32126	0.000	6.905999	0.000	4.852043	0.000	-2.93401	0.000
employed_2013	6.847863	0.000	7.306764	0.000	10.9184	0.000	12.81515	0.000
geo_type_Urban_dum_Bus	-11.65657	0.000	1.483531	0.000	12.30492	0.000	11.04806	0.000

geo_type_Rural_dum_Bus	-3.578325	0.000	8.796318	0.000	16.42844	0.000	11.8862	0.000
_cons	50.34962	0.000	212.9641	0.000	199.2832	0.000	221.2781	0.000
Number of obs =	494.402		474.878		425.159		420.354	
F(10, 2197342) =	17,892.61		10,502.60		10,128.68		8,207.75	
Prob > F =	0.0000		0.0000		0.0000		0.0000	
R-squared =	0.3200		0.2233		0.2365		0.2025	
Adj R-squared =	0.3199		0.2233		0.2365		0.2024	
Root MSE =	102.65		93.072		87.04		84.931	

Source: authors' construction based on IRP5_Merge_2017 dataset.

6 Conclusion

Employment duration and employee churn are very important, as they cost companies money, result in lower employee productivity, and generally influence employee morale. These issues have often been overlooked by past studies.

This research aimed to contribute to our understanding of the impact of accessibility on employment duration. This is in response to the government's focus on transport infrastructure investment and the possible third-order effects of transport—i.e. firm productivity and possible higher incomes and employment—as a policy tool to support development.

Obtaining travel data on these issues is always a problem in developing countries, including South Africa. Researchers have travel data only at city level (for modelling) and country level (NHTS). These data, however, only provide travel information and whether a person is employed or not, which does not reveal employment duration or churn.

A unique dataset of tax administrative records was used for this research. It provided the employment duration for approximately 13.7 million individuals in South Africa in 2017. We were able to aggregate this information by main place level. This dataset, however, has limitations too; with origin (where individuals live) and destination (where individuals work) being poorly populated.

Spatial mapping with these data indicated the job–housing imbalance in South Africa, with workers residing far from job opportunities. The data also indicated this imbalance for different metropolitan areas in South Africa, with Gauteng commuters travelling the furthest. The mean airline distance in urban areas (represented by towns in South Africa) is 10 km, with very few trips exceeding 10 km. These figures indicate that different transport interventions and policies should be considered for different geographical areas, as the demand for certain modes may vary significantly.

Aggregate statistics do not bring out sharp differences across the income distribution in the relationship between airline distance to employment and employment duration. However, income quartiles reveal a positive relationship between duration and distance at higher income levels (income quartiles 3 and 4) and a negative relationship at lower income levels (income quartiles 1 and 2). These variable results can partially be explained by the costs of commuting and the benefit/value of a better salary, which define the employment search space for different income groups. The poor interconnectivity of public transport results in unacceptably high travel

time/costs, which lead to a small employment search space. The converse is true for higher-income employment, where the private transport mode offers a high level of flexibility and access to better employment alternatives.

The results confirm that employment duration, specifically for lower-income workers, is affected by travel distances and that spatial policies should be reviewed. The government should prioritize investments in public transport to provide better transport accessibility to lower-income workers, so as to reduce the income percentage that these individuals spend on transportation costs. Policies should be amended to encourage land use planning that involves allocating housing for lower-income households closer to work opportunities, and in areas that are more accessible by public transport. This will allow a wider choice and improved labour market participation rates. Urban development should focus on existing public transport facilities and routes that do not promote further urban sprawl.

These policy changes have the potential to reduce inequality in the South African workforce by increasing the continuous employment duration of lower-income individuals who are reliant on public transport to commute to and from work, ultimately reducing unemployment through more accessible and affordable transport.

The research should be taken forward by considering possible changes over years, as well as the inclusion of additional variables such as public transport network and transport costs/time as opposed to straight-line distance. The results do suffer from the MAUP, which could be improved if more detailed data were made available at a finer spatial grain for future research.

References

- Bassanini, A., and A. Garnero (2012). 'Dismissal Protection and Worker Flows in OECD Countries: Evidence from Cross-Country/Cross-Industry Data'. IZA Discussion Paper 6535. Bonn: Institute for the Study of Labour (IZA). Available at: <http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=03CF8C1A847987BA38F12318AB6877E5?doi=10.1.1.647.495&rep=rep1&type=pdf> (accessed 20 September 2017).
- Blumenberg, E., and G. Pierce (2014). 'A Driving Factor in Mobility? Transportation's Role in Connecting Subsidized Housing and Employment Outcomes in the Moving to Opportunity (MTO) Program'. *Journal of the American Planning Association*, 80(1): 52–66.
- Brophy, T., N. Branson, R.C. Daniels, M. Leibbrandt, C. Mlatsheni, and I. Woolard (2018). 'National Income Dynamics Study Panel User Manual: Version 1'. Cape Town: South Africa Labour and Development Research Unit (SALDRU), University of Cape Town. Available at: <https://www.datafirst.uct.ac.za/dataportal/index.php/catalog/452/download/9653> (accessed 20 April 2020).
- Department of Transport. (2015). 'National Land Transport Strategic Framework: Final Draft—March 2015'. Available at: https://www.gov.za/sites/default/files/gcis_document/201508/39092gen823.pdf (accessed 16 September 2019).
- Ebrahim, A., and C. Axelson (2019). 'The Creation of an Individual Panel using Administrative Tax Microdata in South Africa'. WIDER Working Paper 2019/27. Helsinki: UNU-WIDER.
- Franklin, S. (2015). 'Location, Search Costs and Youth Unemployment: A Randomized Trial of Transport Subsidies in Ethiopia'. CSAE Working Paper. Oxford: Centre for the Study of African Economies (CSAE).

- Grengs, J. (2010). 'Job Accessibility and the Modal Mismatch in Detroit'. *Journal of Transport Geography*, 18(1): 42–54.
- Johnson, D., M. Ercolani, and P. Mackie (2017). 'Econometric Analysis of the Link between Public Transport Accessibility and Employment'. *Transport Policy*, 60: 1–9.
- Kain, J.F. (1968). 'Housing Segregation, Negro Employment, and Metropolitan Decentralization'. *Quarterly Journal of Economics*, 82(2): 175–97.
- Kerr, A. (2015). 'Tax(i)ing the Poor? Commuting Costs in South Africa'. Working Paper 156. Cape Town: SALDRU, University of Cape Town. Available at: http://opensaldru.uct.ac.za/bitstream/handle/11090/789/2015_156_Saldruwp.pdf?sequence=1 (accessed 14 September 2017).
- Kerr, A. (2017). 'Job Flows, Worker Flows, and Churning in South Africa'. *South African Journal of Economics*, 86(S1): 141–66.
- Kerr, A., M. Wittenberg, and J. Arrow (2013). 'Job Creation and Destruction in South Africa'. Available at: http://www.opensaldru.uct.ac.za/bitstream/handle/11090/608/2013_92.pdf?sequence=1 (accessed 7 September 2017).
- Kgwedi, R., and S.C. Krygsman (2019). 'Accessibility of Public Transport with Focus on Access, Egress and the Line Haul Travel Tie'. Paper presented at the South African Annual Transport Conference (SATC), Pretoria, 10 July.
- Krygsman, S.C., T. de Jong, and O. Verkoren (2016). 'Cape Town and Its Employment Centres: Monocentric, Polycentric or Somewhere In-Between?' In R. Donaldson, G. Visser, J. Kemp, and J. de Waal (eds), *Proceedings of the Centenary Conference of the Society of South African Geographers*. Stellenbosch: Society of South African Geographers.
- Lau Cho-Yam, J. (2010). 'Public Transport and Job-Seeking Range of the Poor in Older Urban Districts in Hong Kong'. *Habitat International*, 34(4): 406–13.
- Melhorado, A.M.C., H. Demirel, M. Kompil, E. Navajas, and P. Christidis (2016). 'The Impact of Measuring Internal Travel Distances on Self-Potentials and Accessibility'. *European Journal of Transport and Infrastructure*, 16(2): 300–18.
- Ngarachu, M., A. Schimmelpfennig, and V. Schöer (2015). 'The Costly Road to Work ? Wages and Transport Costs in South Africa'. IMF Working Paper. Washington, DC: International Monetary Fund (IMF).
- Openshaw, S. (1983). *The Modifiable Areal Unit Problem*. Norwick: Geo Books.
- Phillips, D.C. (2014). 'Getting to Work: Experimental Evidence on Job Search and Transportation Costs'. *Labour Economics*, 29: 72–82.
- Pieterse, D., E. Gavin, and C.F. Kreuser (2018). 'Introduction to the South African Revenue Service and National Treasury Firm-Level Panel'. *South African Journal of Economics*, 86(S1): 6–39.
- Rospabe, S., and H. Selod (2006). 'Does City Structure Cause Unemployment? The Case of Cape Town'. In H. Bhorat and R. Kanbur (eds), *Poverty and Policy in Post-Apartheid South Africa*. Cape Town: HSRC Press.
- Sinclair-Smith, K., and I. Turok (2012). 'The Changing Spatial Economy of Cities: An Exploratory Analysis of Cape Town'. *Development Southern Africa*, 29(3):391–417.
- Statistics South Africa (Stats SA) (2013). *National Household Travel Survey 2013*. Statistical Release P0320. Pretoria: Stats SA. Available at: <http://www.datafirst.uct.ac.za/dataportal/index.php/catalog/501> (accessed 31 July 2019).

- Stats SA (2015). *Measuring Household Expenditure on Public Transport: In-Depth Analysis of the National Household Travel Survey 2013 Data*. Technical report 03-20-11. Pretoria: Stats SA.
- Stats SA (2017a). *Living Conditions of Households in South Africa*. Statistical Release P0310. Pretoria: Stats SA.
- Stats SA (2017b). *Quarterly Labour Force Survey Q1: 2017*. Pretoria: Stats SA. Available at: <https://www.statssa.gov.za/publications/P0211/P02111stQuarter2017.pdf> (accessed 16 October 2019).
- Stats SA (2019a). *Quarterly Labour Force Survey Q1: 2019*. Pretoria: Stats SA. Available at: http://www.statssa.gov.za/publications/P0211/Presentation_QLFS_Q1_2019.pdf (accessed 19 June 2019)
- Stats SA (2019b). *Quarterly Labour Force Survey Q2: 2019*. Pretoria: Stats SA. Available at: <http://www.statssa.gov.za/publications/P0211/P02112ndQuarter2019.pdf> (accessed 16 October 2019).
- Stats SA (2019c). *Quarterly Labour Force Survey Q4: 2019*. Pretoria: Stats SA (producer). Cape Town: DataFirst (distributor).
- Venables, A.J. (2016). *Incorporating Wider Economic Impacts within Cost–Benefit Appraisal*. Paris: Organisation for Economic Co-operation and Development.

Appendix A: 2013–17 SARS IRP5—origin-destination (O-D) dataset variables

The master dataset comprises the cleaned IRP5 2013–17 datasets ('IRP5_2017_cleaned.dta', version 0.6), including the merged variables from the ID panel (version 'Individual_panel_2019_1'), income panel (version 'Individual_panel_2019_1'), and the firm-level panel dataset (version 'FINAL_Panel_01 Aug19'). The dataset consists of 31 columns. The data were accessed between 6 August 2019 and 10 October 2019.

Variable	Example	Description
<i>Identifiers</i>		
id_d	abcdek	Derived variable from ID panel. Identifies unique individuals over all tax years. Referred to as 'employees'.
irp5it3aid	1285012967	Unique IRP5 ID number allocated by SARS after a return is submitted. Indication of the number of jobs per employer or firm.
PAYE_RefNo	DJFKVMDL	The PAYE reference number of the IRP5 certificate. Referred to as 'employers'.
taxrefno	IEOLGMCS	Anonymized unique SARS tax reference number. Variable linked to the firm-level panel.
taxyear	2017	Refers to the tax year from 1 March to the last day of February (i.e. 2017 tax year runs from 1 March 2016 to 28 February 2017).
<i>Demographics</i>		
Age	33	Derived from the date-of-birth variable. Age of the employee within the specific tax year.
Male	1	1 if person is male, 0 if female.
Nationality	1= Foreign, 0= ZA citizen	Derived variable indicating whether the employee is a South African citizen or not. Derived from whether an IRP5 observation is linked to an anonymized South African ID number or an anonymized passport number.
<i>Employment</i>		
employment_start	2016/03/01	Start date of employment in particular tax year.
employment_end	2017/02/28	End date of employment in particular tax year.
employment duration_days	365	Derived variable from employment_start and employment_end variables.
irp5_empl3601	5	Number of IRP5 forms receiving an income 3601 code per firm in tax year.
employed_2016	1	Derived variable: 1 if the person was employed by that company in 2016, 0 if not.
employed_2015	1	Derived variable: 1 if the person was employed by that company in 2015, 0 if not.
employed_2014	1	Derived variable: 1 if the person was employed by that company in 2014, 0 if not.
employed_2013	1	Derived variable: 1 if the person was employed by that company in 2013, 0 if not.
id_d_Branch	4	Derived variable: Number of employees (id_d's) per employer (per PAYE reference number).
irp5_industry_code	MINING AND QUARRYING	Industry code derived from IRP5 form.
Gross_income_d	50,293	From income panel. Sum of all income (business, normal, allowances, fringe benefits, lump sum income, etc.) for an individual within the specific tax year.

Nature_of_person	INDIVIDUAL	Only certificates on behalf of individuals are included. Other options represented in the IRP5 datasets include clubs, partnerships, retirement funds, and associations.
<i>Location</i>		
province_geo	Western Cape	Province in which employee resides. Derived from postal address code.
districtmunicip_geo	City of Cape Town	District municipality in which employee resides. Derived from postal address code.
localmunicip_geo	City of Cape Town	Local municipality in which employee resides. Derived from postal address code.
mainplace_geo	Cape Metro	Main place in which employee resides. Derived from postal address code.
busprov_geo	Western Cape	Province in which the employer (PAYE_RefNo) operates. Derived from postal address code.
busdistmuni_geo	City of Cape Town	District municipality in which the employer (PAYE_RefNo) operates. Derived from postal address code.
buslocmuni_geo	City of Cape Town	Local municipality in which the employer (PAYE_RefNo) operates. Derived from postal address code.
busmainplc_geo	Epping Industria	Main place in which the employer (PAYE_RefNo) operates. Derived from postal address code.
geo_type_Metro_dum_Bus	1	Derived from busmainplc_geo variable: geo-classification of main place according to the 2011 Census survey.
geo_type_Urban_dum_Bus	1	Derived from busmainplc_geo variable: geo-classification of main place according to the 2011 Census survey.
geo_type_Rural_dum_Bus	1	Derived from busmainplc_geo variable: geo-classification of main place according to the 2011 Census survey.
<i>Transport/proximity</i>		
air_dist_km_Internal	50	Airline distance between the main place where the employee resides and the main place of the employer's workplace. This variable includes all internal trips.
air_dist_km	60	Airline distance between the main place where the employee resides and the main place of the employer's workplace, excluding trips made within the same main place.
MP_OD_equal	1	Derived from busmainplc_geo and mainplace_geo variables: 1 if individual lives and works in the same main place, 0 if not.

Source: authors' construction.

Appendix B: SARS IRP5—aggregated at main place level (2017 tax year)

Each row in this dataset ('density_ZA_2017.dta') represents aggregate employment and travel information at main place level in order to conduct spatial mapping. The dataset consists of 1,039 observations and 12 columns.

Variable	Example	Description
MP_CODE	798026	Main place code.
employment	42,016	Derived variable indicating the total number of employees working in the main place.
PAYEnumber_d	1,075	Derived variable indicating the total number of employers located in this main place.
irp5it3aid	50,288	Derived variable indicating the total number of employees residing in this main place.
air_dist_km_Internal	22.39439	Average airline distance for all employees travelling to this main place.
sum_employment_all	1.38e+07	Derived variable indicating the total employment opportunities in the cleaned IRP5 dataset.
sum_employment	1.34e+07	Derived variable indicating the total employment opportunities indicating work address.
sum_irp5it3aid	2,211,554	Derived variable indicating the total employees in the cleaned IRP5 dataset.
employment_w	43,111.77	Derived variable indicating the total number of employees working in the main place weighted according to 'sum_employment_all'.
irp5it3aid_w	312,863.3	Derived variable indicating the total number of employees residing in this main place weighted according to 'sum_employment_all'.
employment_dif	-269,751.5	Derived variable indicating the difference between 'employment_w' and 'irp5it3aid_w'.
Work_Employ_ratio	1.5	Derived variable indicating the ratio between total number of employees in a main place (weighted) and the total number of employees working in the main place (weighted): 'irp5it3aid_w' divided by 'employment_w'.

Source: authors' construction.