

README

South Africa's Fragmented Cities: The Unequal Burden of Labor Market Frictions Javier Baez and Varun Kshirsagar, December 2025

December 16th 2025

Overview

Abstract

Using high-resolution administrative, census, and satellite data, we show that South African cities are characterized by spatial mismatches between where people live and where jobs are located, relative to 20 global peers. Areas within 5 km of commercial centers have 9,300 fewer residents per square km than expected, 60% below the global median. Poor, dense neighborhoods are most affected. In Johannesburg, a 10-percentile increase in distance from the nearest business hub corresponds to a 3.7-percentile drop in asset wealth (a proxy of household wellbeing) and 4.9 percentile drop in employment. In Cape Town, the declines are 4.0 and 3.7 percentiles, respectively. Employment is 87% lower in the poorest decile than the richest in Johannesburg, and 61% lower in Cape Town. These findings suggest that South Africa's spatial organization of people and economic activity constrains agglomeration and reinforces inequality. This methodology provides a scalable and standardized data-driven framework to analyze spatial accessibility and agglomeration frictions in complex, data-constrained urban systems.

Steps to Replicate

This is for the WB Reproducibility Team that has access to the final data

- 1) Unzip the Data file and add it to the folder
- 2) Code/Master has 3 master files that need to be run in order
 - a) run_all1.R (change line 18 to your directory)
 - b) Stata_Master (change line 12)
 - c) Run_all2.R (change line 8)
- 3) The Output will appear in Results (these are the Tables) and Graphs (Maps and Figures). These may be compared against these same folders (Results and Graphs) in the folder Output_This_is_Not_Touched_By_Code

Data Availability

- No data can be made publicly available but the authors have shared all the data with the reproducibility unit to fully reproduce every Table and Figure.

Data Sources

1. **Census Data** contained in Data/Economic/Census_2011

Data format: csv

Please contact: info@statssa.gov.za

The subplace data may be accessed after registration at:
<https://superweb.statssa.gov.za/webapi/jsf/login.xhtml>

This was provided in 2023. See reproducibility notes for the cleaning of the subplaces.

2. **Spatial Tax Data** contained in Data/Economic/Spatial_Tax

Data format: csv

[Explore Tax Data | Spatial Economic Activity Data, South Africa](#)

This was downloaded in 2024

3. **Google Open Buildings** contained in Data/GOB

Data format: parquet

[Open Buildings - Google Research - Open Buildings](#)

This was downloaded in 2025

4. **OSRM Origin-Destination Matrices** contained in Data/OD_Matrices

Data format: csv

The folder contains an R script to download the matrices as well as the meta data needed to replicate these.

This was downloaded in 2025

5. **OSM data on roads and POIS** for South Africa contained in Data/OSM

Data format: gpkg and pbf

[Download OpenStreetMap for South Africa | Geofabrik Download Server](#)

This was downloaded in 2025

6. **GHS Building Volumes** contained in Data/Remote_Sensed

Data format: tif

https://human-settlement.emergency.copernicus.eu/ghs_buV2023.php

Pesaresi M., Politis P. (2023):

GHS-BUILT-V R2023A - GHS built-up volume grids derived from joint assessment of Sentinel2, Landsat, and global DEM data, multitemporal (1975-2030). European Commission, Joint Research Centre (JRC)

PID: <http://data.europa.eu/89h/ab2f107a-03cd-47a3-85e5-139d8ec63283>,
doi:[10.2905/AB2F107A-03CD-47A3-85E5-139D8EC63283](https://doi.org/10.2905/AB2F107A-03CD-47A3-85E5-139D8EC63283)

This was downloaded in 2023

7.GHSL Population Gridded Resolution of 1Km - World Mollweide (EPSG:54009)
contained in Data/Remote_Sensed

Data Format: tif

[Global Human Settlement - GHS-POP_GLOBE_R2023A - European Commission](#)

Dataset:

Schiavina M., Freire S., Carioli A., MacManus K. (2023):

GHS-POP R2023A - GHS population grid multitemporal (1975-2030).European
Commission, Joint Research Centre (JRC)

PID: <http://data.europa.eu/89h/2ff68a52-5b5b-4a22-8f40-c41da8332cfe>,
doi:[10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE](https://doi.org/10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE)

This was downloaded in 2023

8.Night Time Lights contained in Data/Remote_Sensed

Data format: tif

[VIIRS Nighttime Light](#)

C. D. Elvidge, K. Baugh, M. Zhizhin, F. C. Hsu, and T. Ghosh, "VIIRS night-time lights,"
International Journal of Remote Sensing, vol. 38, pp. 5860–5879, 2017.

This was downloaded in 2023

9. Shape Files contained in Data/Shape_Files

Data format: .shp and ancillary

All city and country boundaries are from GADM [GADM](#)

The subplace shape files (SP_SA_2011) are from the South Africa Stats Office, Please
contact: info@statssa.gov.za

The UberH3_7 are from [Explore Tax Data | Spatial Economic Activity Data, South Africa](#)

Statement about Rights

- I certify that the authors of the manuscript have legitimate access to and permission to use the data used in this manuscript.

Instructions for Replicators

New users should follow these steps to run the package successfully: - Users must first have access to all data files if they are not included in the reproducibility package. They should go to the mentioned links, download the listed files, and place them in the data folder. - Update the following files with your directory paths

Code/Master has 3 master files that need to be run in order

- 1) run_all1.R (change line 18 to your directory)
- 2) Stata_Master (change line 12)
- 3) Run_all2.R (change line 8)

List of Exhibits

Figures: With some exception listed below, all Figures and Maps produced by the code are labelled as in the paper

Except the following

Figure 3a is Figure 3 in the paper

Figure 3b is Figure 4 in the paper

Figure 4 is Figure 5 in the paper¹

Figure 7a: johannesburg_4panel_quintiles.png

Figure 7b: capetown_4panel_quintiles.png

Figure 7c: durban_4panel_quintiles.png

All Tables are labelled as in the paper

Except the following

Table 1 is in Results/Table_1/Table_1_100K_09_2020

The Appendix Table A2 takes the coefficients from cell G12 for all the tables in Results/Table_1/

Appendix_Table_A1 is in Results/Table_1/Table_A1_100K_09_2020

¹ The bar graphs produced by the code are labelled Figure 5a and 5b by the code but are no longer in the paper. There were slight discrepancies, likely due to version differences and we have chosen to omit them from the paper.

Requirements

Software Requirements

List all software requirements, including versions, dependencies, libraries, environment

- **R Version 4.5.1**

The renv folder, Rprofile, renv.lock allows users to get the versions of all the R packages used (including the tidyverse and the geospatial suite of packages)

- **Stata 16.1**

- outreg2, coeffplot

Memory and Runtime and Storage Requirements

The Data folder is 46 gb when unzipped.

Code/Master has 3 master files that need to be run in order (3 clicks)

i)run_all1.R (change line 18 to your directory): Takes about 40 minutes

ii) Stata_Master (change line 12): Takes about 10 minutes

iii)Run_all2.R (change line 8): Takes about 100 minutes

This was run on a machine with 64 GB of RAM and an Intel(R) Core(TM) i9-11900K CPU @ 3.50GHz processor using R 4.5.1 and Stata 16.1

Reproducibility Notes

- 1) **Deterministic volatility and percentiles (Tables 4a–4b).** To minimize cross-platform drift, we made the construction of the low-wage employment percentile and volatility percentile fully deterministic in Stata 16.1. For volatility (file `volatility_hex7_w1.do`), we: (i) enforce double precision, (ii) sort observations by `hex7` and time, (iii) run per-hex seasonal regressions, and (iv) snap the residuals and the collapsed standard deviation to 1e-10 and 1e-6 respectively using `round()`. Subsequent scripts that create the percentile variables for Tables 4a and 4b sort the data and apply `xtile` to these rounded values (`n(100)`), ensuring that percentile breakpoints are deterministic rather than depending on differences across Stata versions.
- 2) **Cleaning the subplace shape files.** There was a reproducibility issue in the first test caused by invalid polygons in the official SP_SA_2011 subplace shapefile (the World Bank Reproducibility team got an `s2/wkt` error regarding open loops, this ran on our desktop using `geos` without any issues) . We diagnosed this using `st_is_valid(reason = TRUE)`, created a cleaned, version (SP_SA_2011_clean.gpkg) where all 22,108 geometries are now valid, and updated the analysis scripts to use

this cleaned layer. We treat SP_SA_2011_clean.gpkg as a data input for the reproducibility package. For transparency, we have included (i) the original SP_SA_2011 shapefile, and (ii) the R cleaning script (clean_SP_SA_2011.R) that reproduces the cleaned GPKG from the original file, so that the cleaning step is fully documented.