README: Spatial Job Accessibility Index

Reproducibility package for “Urban Job Accessibility for the Poor” (Davalos, Kshirsagar, Stemmler), Prosperity Insights series

Version: 10 August 2026.

1. Overview

This reproducibility package contains the code and data needed to replicate all exhibits in “Urban Job Accessibility for the Poor.” The analysis measures spatial disparities in job accessibility across seven cities in four countries: Accra (Ghana); Bogotá and Cali (Colombia); Cape Town, Johannesburg, and eThekwini (South Africa); and Metro Manila (Philippines). Hanoi (Vietnam) is an out-of-sample deployment city: the trained model is applied to it, but it is not part of the training data.

The package is organized into two parts, corresponding to the two sections of the paper. Part 1 loads and processes census and employment microdata, computes car and transit travel-time matrices, constructs the job accessibility measures (cumulative opportunity counts, the gravity-based JAI, and jobs-per-resident JPR), and generates the Part 1 exhibits. Part 2 builds predictive models that classify areas by accessibility typology using satellite-derived features, and produces the diagnostic, targeting, and out-of-sample exhibits.

Replicators should read the Setup and Execution section, set the data folder path in `code/00.master.R`, and run that script end to end. Section 10 states what the verification team can run and what a public replicator can run. The two differ because of data confidentiality.

2. Data Availability

- ☐ All data are publicly available.
- ☒ Some data cannot be made publicly available.
- ☐ No data can be made publicly available.

Confidential population and economic census microdata for Ghana, Colombia, the Philippines, and South Africa were shared under restricted-use agreements with the respective statistical offices and government agencies. These files cannot be redistributed. In addition, no census-derived or welfare-derived material is published in any form: this covers the Part 1 processed datasets and the Part 2 outcome and label files, all of which are derived from the confidential microdata. All confidential inputs have been shared privately with the World Bank reproducibility verification team, which can therefore run the package end to end.

The public release comprises the package code, this README with its Data Availability Statement, the outputs folders, and the processed remote-sensed feature files, the only data files in the public package, released under CC-BY-4.0 (Section 13). All public raw sources (administrative boundaries, GTFS feeds, OpenStreetMap extracts, and published poverty indices) are freely downloadable; URLs and access details are in Section 12.

2.1 Statement about Rights

☒ I certify that the author(s) of the manuscript have legitimate access to and permission to use the data used in this manuscript.

☐ I certify that the author(s) of the manuscript have documented permission to redistribute/publish the data contained within this replication package.

Redistribution permission applies only to the processed remote-sensed feature files (released under CC-BY-4.0); the remaining data are not redistributed.

3. Verification status

All exhibits reproduce. Four of the seven Part 1 processed datasets were verified byte-identical (same MD5 checksum) against the originals, covering each of the four distinct processing paths (Ghana, Colombia, South Africa, Philippines); the remaining three cities each share a verified code path and were verified at the exhibit level. In Part 2, the main results table (Table 2), the typology recall table (Table A12b), and the Hanoi out-of-sample scores are byte-identical.

Artifact	Produced by	Result
Processed datasets: Accra, Cali, Johannesburg, Metro Manila	`part1_02a / part1_02b`	Byte-identical
Figure 1, all 7 city panels	`part1_05`	Matches report
Figure A2, all 7 cities	`part1_05`	Matches report
Table 2, main results	`part2_02a_engine_dc.R`	Byte-identical
Table A5, Moran's I	`part2_03a_diagnostics_dc.R`	20/20 rows match
Table A12b, typology recall (all four engines plus random baseline)	`part2_05_fig_typology_recall.R`	Byte-identical
Hanoi out-of-sample scores	`part2_09_predict_hanoi.R`	Byte-identical

Byte-identity of Table 2 is a strong determinism check: it contains bootstrap confidence bounds from 1,000 spatial-block resamples computed on a parallel cluster, so bit-for-bit reproduction confirms that the seed policy in `part2_config.R` (L'Ecuyer-CMRG RNG, fixed seed, per-worker RNG streams, single-threaded BLAS and data.table, C locale) works as designed. Verifiers are welcome to confirm checksums with `md5sum`.

4. List of Exhibits

The provided code reproduces:

☒ All numbers provided in text in the paper

☒ All tables and figures in the paper

☐ Selected tables and figures in the paper, as explained and justified below

Tables 1, A1a, A1b, and A3 in the report are descriptive or definitional (proposed methodologies, data-source inventories, and ablation scenario definitions) and are not produced by code; every computed exhibit is listed below.

4.1 Part 1 Figures

All produced by `part1_05.visualizations.R`; saved to `outputs/figures/`.

Exhibit	Output filename	Note
Figure 1	`fig1_jobs_by_travel_time_band_all_cities.png`	Jobs reachable by travel-time band, all 7 cities
Figure 2a	`fig2_cross_city_jpr_comparison.png`	Cross-city JPR comparison
Figure 2b	`fig2_cross_city_jai_comparison.png`	Cross-city JAI comparison
Figure 3	`fig3_accra_formal_informal_jpr_pct_diff.png`	Accra formal vs informal JPR
Figure 4	`fig4_{city}_sector_jpr_pct_diff.png`	Sector JPR; Accra, Bogotá, Johannesburg, Metro Manila. Filenames as generated by the code; the Bogotá file is fig4_bogot__sector_jpr_pct_diff.png (non-ASCII characters are replaced in filenames)
Figure 5	`fig5_simulation_jai_impact.png`	Calls run_simulation_and_plot() from part1_04b.sim_tt_fct.R
Figure 6a	`fig6a_accra_transit_coverage_map.png`	Accra poverty and employment density with transit stops (reads gha/accra.gtfs/stops.txt on every run)
Figure 6b	`fig6b_accra_transit_walk_comparison.png`	Accra walking vs public transport job access
Figure 6c	`fig6c_accra_transit_jobs_by_time_threshold.png`	Accra transit jobs by time threshold
Figure 7a	`fig7a_accra__transit__jai_jpr_pct_diff_panel.png`	Accra transit JAI and JPR pct difference
Figure 7b	`fig7b_accra__transit__formal_informal_jpr_pct_diff.png`	Accra transit formal vs informal JPR
Figure 8a	`fig8_accra_bivariate_jpr.png`	Accra bivariate map: poverty × JPR (car)
Figure 8b	`fig8_accra__transit__bivariate_jpr.png`	Accra bivariate map: poverty × JPR (transit)

4.2 Part 2 Figures and Tables

Saved to `Outputs/predictions/` in the subfolders shown (note the capital O; see Section 8).

Exhibit	Output filename	Script
Figure 9	`Diagnostics_DC/fig_roc_dc_rep.png`	`part2_03a_diagnostics_dc.R`
Figure 10	`Diagnostics_DC/fig_map_dc_appendix_b.png`	`part2_03a_diagnostics_dc.R`
Figure 11	`Targeting_Profile/fig_composition_bars.png`	`part2_07_fig_targeting_profile.R`
Figure 12a	`Diagnostics_DC/panel_dc_Bogota_rep.png`	`part2_03a_diagnostics_dc.R`
Figure 12b	`Feature_Maps/feature_maps_Bogota.png`	`part2_08_fig_feature_maps.R`
Table 2	`Stage2_DC/09a_results_table.csv`	`part2_02a_engine_dc.R`

4.3 Annex Figures and Tables

Exhibit	Output filename	Script / note
Figure A1	`figa1_{city}_population_density / _employment / _poverty_rate.png`	part1_05; 3 files × 7 cities
Figure A2	`figa2_fig_jobs_per_wap_by_quintile.png`	part1_05; all cities
Figure A3	`figa3_{city}_jpr_index.png`	part1_05; 7 files
Figure A4	`figa4_{city}_jai_pct_diff.png`	part1_05; 7 files
Figure A5	`figa5_{city}_sector_jpr_pct_diff.png`	part1_05; Cape Town and eThekwini
Figure A6	`Label_Map/fig_label_map.png`	part2_06_fig_label_map.R

Exhibit	Output filename	Script / note
Figure A7	`Diagnostics_DC/fig_boxplot_dc.png`	part2_03a
Figure A8	`Diagnostics_DC/fig_appendix_ols_shap_dc.png`	part2_03a
Figure A9	`Diagnostics_DC/fig_confusion_features_dc.png`	part2_03a
Figure A10	`Diagnostics_DC/fig_corr_dc_rep.png`	part2_03a
Figure A11	`Ablation/fig_ablation_recall_5bar.png`	part2_04; the 5th bar is the RANDOM baseline
Figure A12a	`Diagnostics_PD/fig_score_by_3group_m2_pd.png`	part2_03b
Figure A12b	`Typology_Recall/fig_typology_recall.png`	part2_05
Figures A13–A14	`Hanoi_Predictions/Hanoi_final_A2_M2_*.png`	part2_09; feature maps, priority and uncertainty
Table A2	`table_a2_cities.md, table_a2_cities.docx; raw values in table_a2_values.csv`	part1_04c.city_table.R + part1_05; uniform definition, FUA shown as N.A. (see Section 11.1)
Tables A4a, A4b	`Diagnostics_DC/tbl_gbm_hyperparams_dc.csv, tbl_rf_hyperparams_dc.csv`	part2_03a
Table A5	`Diagnostics_DC/tbl_moran_i_d.csv`	part2_03a
Table A6a	`Diagnostics_DC/tbl_feature_comparison_dc_long.csv`	part2_03a; long format
Table A6b	`Diagnostics_DC/tbl_feature_comparison_dc_wide.csv`	part2_03a; wide format

5. Requirements

5.1 Computational requirements

Component	Requirement	Note
Operating system	Windows 11	The package was built and verified on Windows.
R	4.5.1	Some packages warn “built under R 4.5.3”; patch releases within 4.5.x are binary-compatible and the warning is benign.
renv	1.2.3	Restores the project library from the lockfile at the project ROOT (the original README said code/renv.lock; the root location is correct).
GDAL / PROJ / GEOS	3.12.1 / 9.7.1 / 3.14.1	System libraries reported by sf at startup; renv cannot pin these, so they are recorded here. The exhibits were produced with these versions.
Java 21 (Azul Zulu OpenJDK recommended)	NOT required for reproduction	Needed only if a travel-time cache file is deleted, which re-triggers the Accra r5r transit/walk routing. With the caches present, as shipped, no Java is needed.
OSRM	NOT required for reproduction	Car travel-time matrices ship pre-computed. OSRM matters only if the frozen routing workflows are re-run; the version used originally was not recorded.

Cores and parallel RNG. `part2_config.R` (line 164) sets `N_CORES <- min(6L, max(1L, parallel::detectCores() - 2L))`, and cluster size determines RNG stream allocation via `clusterSetRNGStream`. The verified runs used a 6-worker cluster (the value N_CORES resolves to on any machine with 8 or more cores). A machine with fewer than 8 cores resolves to a smaller cluster and is the most plausible source of cross-machine divergence in Part 2.

5.2 Memory, runtime, and storage

Storage: approximately 70 GB for the input data (private + public + predictions), approximately 160 MB for the Part 1 processed data, and approximately 240 MB for outputs. Runtime: a full run of the pipeline (raw data through Part 1 and Part 2, with all travel-time caches present) takes approximately 1 hour. The frozen preparation workflows, which a replicator does not run, take approximately 3 additional hours of compute when executed with all inputs in place (Section 10). The original README's figure of approximately 8 hours reflected running everything, including routing, on the original hardware.

6. R environment (renv)

All R package dependencies are managed via renv and recorded in `renv.lock` at the project root. Running `00.master.R` installs renv if absent and restores all packages automatically via `renv::restore()`; on success renv reports that the library is synchronized with the lockfile. Key packages include sf, ggplot2, patchwork, tidyverse, foreach, and fixest.

6.1 Lockfile provenance and modifications

The lockfile is based on the one used to produce the report, preserved to keep fidelity with the report rather than regenerated. Two sets of modifications were required to make it restorable on a clean Windows machine. Because renv::restore() is transactional, a single failing package aborts and rolls back the entire restore; the packages below made restore fail while contributing nothing to the runnable pipeline.

Packages removed (10)

Package(s)	Rationale
`duckdb, duckdbfs`	Used nowhere in the runnable pipeline. The only reference was a stray library(duckdbfs) call in part1_01.setup.R (now commented out). Real usage is confined to the frozen prep script buildings_download.R (Section 11).
`mapview, leafem, leaflet, leaflet.providers, leafpop, satellite, raster, sp`	mapview is loaded once in setup (now commented out) and never used; no mapview(), leaflet(), raster() or sp:: call exists in any of the 28 scripts. sp could not be built from source on the target system, and removing this dependency chain was required for restore to complete.

Package versions re-recorded (8)

These versions were archived on CRAN with no Windows binary available, and their source builds failed; each was re-pinned to the nearest available version. The byte-identical verification results (Section 3) confirm that none of these substitutions changed any output; fixest is the only estimation package among them.

Package	Original lockfile	Now pinned
`Rcpp`	1.1.1-1.1	1.1.2
`rlang`	1.2.0	1.3.0
`xfun`	0.58	0.60
`zip`	3.0.0	3.0.1
`selectr`	0.5-1	0.6-0
`xml2`	1.5.2	1.6.0

Package	Original lockfile	Now pinned
fixest	0.14.1	0.14.2
terra	1.9-27	1.9-34

In addition, `renv/settings.json` (shipped with the package) records `ignored.packages = c("r5r", "rJava", "duckdb", "duckdbfs", "mapview")`. r5r remains in the lockfile because the code scan detects it inside the cache-guarded Accra routing block; rJava was de-scoped from setup so that no Java installation is required unless that guarded block actually runs. If library(rJava) fails when re-routing, set JAVA_HOME at the top of 00.master.R.

6.2 Two configuration traps

Workspace restore. Disable "Restore .RData into workspace at startup" in RStudio (Tools → Global Options). With it enabled, objects from a previous session are silently injected into a "fresh" run, a reproducibility hazard.

unzip masking. `library(zip)` (loaded by setup) masks `utils::unzip` with a different signature, so a bare `unzip(..., list = TRUE)` fails with "unused argument". Call `utils::unzip()` explicitly; the Ghana GTFS step uses `zip::unzip()` and is unaffected.

7. Folder structure

```
<PROJECT ROOT>  (open the .Rproj here)
|-- .Rprofile                    (activates renv)
|-- renv.lock                    (lockfile -- at the ROOT)
|-- renv/                        (activate.R, settings.json, library/)
|-- Job_Accessibility_Index.Rproj
|-- code/
|   |-- 00.master.R              (entry point -- set data_dir, line 11)
|   |-- part1_01 ... part1_05    (Part 1 scripts; incl. new part1_04c)
|   |-- part2_config.R, part2_01 ... part2_09   (Part 2: 14 scripts)
|   `-- Data_Prep_Code/          (frozen upstream prep -- documentation only)
|-- data/processed/{col,gha,phl,zaf}/   (Part 1 processed data)
|-- 2.data/                      (raw data folder = data_dir; see Section 12)
|   |-- private/{gha,col,phl,zaf}/
|   |-- public/{gha,col}/ + World Bank Official Boundaries - Admin 2.gpkg
|   `-- predictions/Data/        (Part 2 inputs: Features, Outcomes, ...)
|-- outputs/                     (Part 1: figures/, tables/)
`-- Outputs/predictions/         (Part 2; 13 subfolders, auto-created)
```

Two output trees are intentional and not unified: Part 1 writes to `outputs/` and Part 2 writes to `Outputs/predictions/` (the original README's `outputs/predictions/` was a typo). The prep code under `code/Data_Prep_Code/` documents how the frozen inputs were produced and is not executed by the master script. The South Africa and Philippines preparation code remains bundled with the raw data it processes, in `private/zaf/South_Africa_Data_Prep/` and `private/phl/Philippines_Data_Prep/`, each with its own README, as originally delivered to the verification team.

8. Setup and Execution

1. Install R 4.5.1 from CRAN (Windows).
2. Open `Job_Accessibility_Index.Rproj` in RStudio (this anchors `here()` and activates renv). Disable .RData restore (Section 6.2).
3. Set `data_dir` at line 11 of `code/00.master.R` to your local data folder, which must contain `private/`, `public/`, and `predictions/`.
4. Run `code/00.master.R` top to bottom. It validates the data folder, installs renv if needed, restores packages, and runs the full pipeline from Part 1 raw data through Part 2, including the Hanoi deployment.

9. Scope: what verifiers and public replicators can run

Verification team. The reproducibility verification team receives the complete data folder privately, including all confidential inputs. From that starting point the package runs end to end from raw data: both Part 1 processing paths, all exhibits, all four Part 2 engines, and the Hanoi deployment.

Public replicators. The confidential microdata cannot be redistributed, and no census-derived or welfare-derived material is published in any form (Section 2). The public package therefore contains the code, this README and DAS, the outputs, and the processed remote-sensed feature files (CC-BY-4.0). A public replicator can run the Part 2 scripts whose inputs are limited to those features, reproducing the exhibits that depend only on them; exhibits requiring welfare or census-derived inputs reproduce for the verification team but not for a public user. Public raw sources (Section 12) allow a public user to rebuild the remote-sensed inputs from scratch, subject to the portability caveats in Section 11.

Script by script, the position is as follows. From the public package, `part2_08_fig_feature_maps.R` runs directly: its only inputs are the shipped feature files. Every other script requires confidential material at some stage: in Part 1, the `02a`/`02b` scripts read the confidential microdata and the downstream scripts read the processed datasets derived from it; in Part 2, `part2_01_labels.R` reads the confidential outcomes, the four engines, the label map (`part2_06`), and the Hanoi deployment (`part2_09`, which retrains on the seven cities' labels) read the labels derived from them, and the diagnostics and remaining figure scripts (`part2_03a/b`, `part2_04`, `part2_05`, `part2_07`) consume model outputs produced by the engines.

10. Pre-computed (frozen) workflows

Four upstream preparation workflows are shipped as frozen outputs plus documented code, rather than as steps a replicator re-executes. The reason is portability: OSRM and r5r travel times vary with the routing-engine build and the underlying OpenStreetMap extract; several steps download from sources that change over time (OSM, VIDA building footprints); and the spatial stack (sf → GDAL/PROJ/GEOS) is sensitive to system library versions that renv cannot pin. Re-running these steps would produce small differences that are not analytically meaningful but would break exact reproduction of the published exhibits. Making them portable would require full containerization, which is out of scope for this package.

All four workflows appear in `00.master.R` as commented-out blocks with these notes inline. The one routing step handled differently is the Accra r5r transit/walk routing inside `part1_02a`: it is cache-guarded rather than commented out, executing only if its travel-time cache files are absent. Deleting a cached matrix forces

recomputation from scratch (requiring Java 21), and results may differ marginally if the OSM extract or GTFS feed has changed.

Paths inside the prep code. Every script in `Data_Prep_Code/` resolves its paths from a single environment variable, `JAI_ROOT`, reading and writing under `<JAI_ROOT>/Data/`. The default value is the original development machine's project root and will not exist elsewhere; in this package's layout the same tree lives at `<data_dir>/predictions`, so anyone re-running a prep script should first set `Sys.setenv(JAI_ROOT = file.path(data_dir, "predictions"))`. External inputs outside the package (country OSM extracts for OSRM, and similar) are likewise configured by environment variables documented in each script's header (`OSM_PBF_ROOT`, `GADM_DIR`, and others). Collectively, the four workflows take approximately 3 hours of compute when executed with all inputs in place; the per-workflow split was not recorded.

Workflow	Code	Key frozen outputs	Last run / runtime
1. South Africa prep	`South_Africa_Data_Prep/` (own README); `od_matrix_southafrica.R` in `Data_Prep_Code/`	Cleaned subplace boundaries; three city subplace lookups; three OSRM car OD matrices; SEAD-SA employment toolkit (restricted tax data; hex7 aggregate is the shippable form)	input access dates 07/2023–12/2025; OSRM version unrecovered; runtime not recorded
2. Philippines prep	`Philippines_Data_Prep/` (own README); `od_matrix_philippines.R` in `Data_Prep_Code/`	Harmonized barangay sf object; OSRM OD matrix (Metro Manila). Only the dated output folder JAI_Harmonized_PH_4_28_26/ is read; the undated folder is superseded	On 28 April 2026 (inferred from the dated folder name); inputs accessed 11/2025; runtime not recorded
3. Part 2 data prep	`Data_Prep_Code/` (master: `run_data_prep.R`)	Per-city Outcomes (confidential), the nine remote-sensed Features, OD matrices, road layers	input access dates 01/2026 (VIDA, GHSL, OSM) and 07/2023 (NTL); runtime not recorded
4. Hanoi feature prep	`Data_Prep_Code/Hanoi/`	Hanoi_features.rds, the only Hanoi artifact the pipeline reads; the eight Hanoi/ subfolders are intermediates	Vietnam OSM extract accessed 01/2026. The deployment script part2_09 is NOT frozen: it runs in the pipeline and reproduces the published scores byte-identically

10.1 Part 2 prep script inventory (Data_Prep_Code/, not executed by the master)

Script	Purpose
`run_data_prep.R`	Master script for Part 2 data preparation
`buildings_download.R`	Downloads VIDA building footprints via DuckDB httpfs (sole user of duckdb/duckdbfs)
`morphology_features.R`	Building morphology features from footprints
`raster_features.R`	GHSL volume, height, and nighttime-lights features
`spatial_features.R`	Spatial predictor features
`neighborhoods.R`	Neighborhood units for feature aggregation
`boundaries_clean.R`	City boundary cleaning and standardization
`roads_extract.R`	Road networks from country OSM PBF files
`od_matrix_colombia.R / _ghana.R / _southafrica.R / _philippines.R`	OSRM OD travel-time matrices
`outcomes_colombia_accra.R / outcomes_southafrica.R / outcomes_ph_ncr.R`	JAI label inputs (confidential outcomes)

Script	Purpose
Hanoi/	Hanoi out-of-sample deployment data prep

10.2 Running the Hanoi deployment as a standalone workflow

This section is written so that a third party can rebuild the Hanoi features from public data and score the city without help from the authors. The deployment has two stages. Stage one, feature construction, uses only public data and the eight scripts in `code/Data_Prep_Code/Hanoi/` (that folder's own README.md documents the same flow). Stage two, scoring, is the pipeline script `part2_09_predict_hanoi.R` and has one access condition, stated at the end of this section.

Setup

Open the project and restore the environment as in Section 8. Then set one environment variable so the prep scripts resolve against the package data folder: `Sys.setenv(JAI_ROOT = file.path(data_dir, "predictions"))`. Two optional variables cover external inputs: `GADM_DIR` points to a folder holding the GADM Vietnam files, and `GHSL_FUA_PATH` overrides the location of the GHS functional urban areas layer if it is not in `predictions/Data/GHSL_FUA/`. One extra dependency applies to the download step only: `02b_download_gob_hanoi.R` uses the duckdb and duckdbfs packages, which are not in the lockfile because nothing in the pipeline needs them; install them separately if you run that step.

Inputs (all public)

Input	Where to get it	Used by
GHS-FUA functional urban areas layer (GHS_FUA_UCDB2015 R2019A)	Ships in predictions/Data/GHSL_FUA/. Source: EC JRC, CC-BY 4.0.	`01b`
GADM Vietnam administrative boundaries (levels 2 and 3)	gadm.org, free download. Place in the folder named by GADM_DIR.	`01b, 03b`
VIDA Google-Microsoft-OSM Open Buildings, Vietnam	Downloaded over the network by the script itself (DuckDB httpfs from Source Cooperative). No manual download needed.	`02b`
GHSL building volume and height rasters, nighttime lights, GHS-POP	Ship in predictions/Data/Rasters/. Sources and licenses in Section 12.5.	`06b, 08a`

Run order

Run the eight scripts in this order (the numbering has a gap; there is no 05b). Each writes its outputs under predictions/Data/Hanoi/, except the last, which writes the final artifact to predictions/Data/Features/.

Step	Script	Reads	Writes
1	`01b_hanoi_boundary_ghsl.R`	GHS-FUA layer, GADM level 2	Hanoi core and buffered boundaries (Clean_Boundaries/)
2	`02b_download_gob_hanoi.R`	Boundaries; building footprints over the network	GOB parquet files and download manifest
3	`03b_hanoi_neighborhoods.R`	GADM level 3, buffered boundary	Hanoi_neighborhoods.gpkg
4	`04b_gob_features_hanoi.R`	Building footprints, neighborhoods	GOB_Features/ (morphology features)
5	`06b_raster_features_hanoi.R`	Shared rasters, neighborhoods	Raster_Features/
6	`07b_spatial_features_hanoi.R`	Raster features	Spatial_Features/

Step	Script	Reads	Writes
7	08a_population_hanoi.R	GHS-POP raster, spatial features	Final_Features/Hanoi_all_features.gpkg and .csv
8	09b_hanoi_artifact.R	Hanoi_all_features.gpkg	Features/Hanoi_features.rds (the artifact the pipeline reads)

The final artifact has a fixed contract, enforced by step 8: `unit_id` (derived from GADM `GID_3`), the key and label fields, `eligible`, `nbhd_area_km2`, `pop_total`, `pop_density_km2`, the nine raw features, `lon` and `lat` (WGS84 centroid degrees), and WGS84 geometry.

Checking your results, and scoring

Every intermediate of this workflow ships in predictions/Data/Hanoi/, so each step's output can be compared against the shipped copy as you go. Note that step 2 downloads current building-footprint data, so features rebuilt today can differ marginally from the shipped, frozen versions; for exact reproduction of the published scores, use the shipped intermediates or the shipped final artifact, which was verified byte-identical in July 2026.

Scoring is then one pipeline step: source `code/part2_config.R` and `code/part2_09_predict_hanoi.R`. The access condition: this script trains the model on the seven training cities' Features and Labels before applying it to Hanoi. The Features are public, but the Labels derive from confidential census and welfare data, so scoring runs for the verification team and for holders of the confidential inputs. The published Hanoi scores ship in the outputs for comparison.

10.3 Deploying the model to a new city (worked example: Mombasa)

The Hanoi workflow in Section 10.2 is a template, and this section explains how to apply it to a city outside the package. The example used throughout is Mombasa, Kenya. The eight Hanoi scripts are structured to be city-agnostic, but they were run for Hanoi, so deploying elsewhere is a light adaptation exercise rather than a single command. A user comfortable in R should expect to edit city names, one ISO3 country code, one map projection, and file paths, and nothing else. The model itself requires no retraining decisions from the user: it consumes the same nine features, built the same way, for any city.

What a new city needs

Every input is public and global. Four things must be obtained for the new city; two more are already in the package.

Input	Where to get it for a new city	Notes
City boundary (functional urban area)	Already in the package: the GHS-FUA layer in predictions/Data/GHSL_FUA/ is a global vector, so it contains Mombasa. No download needed.	Used by the step 1 script to cut the core and buffered boundaries.
Neighborhood units (small administrative areas)	GADM (gadm.org, free): download the country file and use the finest level available. For Kenya, wards are the natural unit. If GADM is too coarse for the country, the OCHA Common Operational Datasets (data.humdata.org) are an alternative.	Hanoi used GADM level 3. Whatever layer is chosen supplies the unit identifier that becomes unit_id.
Building footprints	Downloaded by the step 2 script itself. It pulls the VIDA Google-Microsoft-OSM combined footprints from	URL pattern: data.source.coop/vida/google-microsoft-osm-open-buildings/geoparquet/by_country/country_iso={ISO3}/{ISO3}.parquet.

Input	Where to get it for a new city	Notes
	Source Cooperative by ISO3 code. Change the code to KEN and the script fetches the Kenya file over the network.	Requires duckdb and duckdbfs (Section 10.2). Country files can be large; the script clips to the city boundary.
GHSL built-up volume and height rasters	EC JRC, free, CC-BY 4.0. Volume: human-settlement.emergency.copernicus.eu/ghs_buV2023.php (GHS-BUILT-V, epoch 2020, 100 m). Height: .../ghs_buH2023.php (GHS-BUILT-H ANBH, epoch 2018, 3 arcsec). Download the tiles covering the city from the tile grid on each product page.	Use the same product versions and epochs as the training features, listed here, so the new city's features are comparable to the model's training data.
Nighttime lights	Earth Observation Group: eogdata.mines.edu/products/vnl/ (free registration). Use the annual VNL V2 composite for 2019, the vintage the model was trained on.	One global file; place it in predictions/Data/Rasters/NTL/ or point the script at it.
Population raster	EC JRC GHS-POP R2023A, epoch 2020, 100 m, from the same JRC site (DOI 10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE). Download the tiles covering the city.	Supplies pop_total and pop_density_km2 for the artifact.

Download the tiles above and either mosaic them into the same folders or point the scripts at the new files.

Adapting the scripts

Copy `code/Data_Prep_Code/Hanoi/` to a new folder (for example `Mombasa/`) and work through the eight scripts in the Section 10.2 run order, making these changes. In step 1, replace the Hanoi city name with the new city's name as it appears in the GHS-FUA layer, and check visually that the selected polygon is the right one. In step 2, set the ISO3 code (KEN). In step 3, point `GADM_DIR` at the downloaded country boundaries and set the administrative level for the neighborhood units. Throughout, replace the projected coordinate system: the Hanoi scripts compute planar areas and centroids in UTM zone 48N (EPSG:32648), and Mombasa sits in UTM zone 37S (EPSG:32737); use the UTM zone of whatever city is being run. Steps 4 through 7 need only the city-name token changed, since they read whatever the earlier steps wrote. In step 8, set `unit_id` to derive from the identifier field of the chosen admin layer (the Hanoi script uses GADM's `GID_3`). The output naming convention matters: the final artifact must land in `predictions/Data/Features/` as `{City}_features.rds` with the contract columns listed in Section 10.2, because that is where `feature_path()` looks for it.

Scoring the new city

Copy `part2_09_predict_hanoi.R` and change the deployment city token from Hanoi to the new city, so the script reads `feature_path("Mombasa")` and writes to a new output folder. The access condition from Section 10.2 applies unchanged: the script trains on the seven cities' Features and Labels before predicting, and the Labels are confidential, so this step runs for the verification team, for World Bank teams with access to the confidential inputs, or with a fitted model object supplied by the authors. Everything upstream of scoring, which is the bulk of the work, requires no restricted data at all.

A note on effort. The feature construction for one new city is bounded by the building-footprint download and processing; the remaining steps are minutes each on the package's reference hardware. Runtime depends mostly on the size of the country footprint file and the internet connection. For a first deployment it

is sensible to run the eight steps interactively, checking each output against the shipped Hanoi equivalents as a structural reference.

11. Definitions and changes relative to the published report

11.1 Table A2, definition change

The published Table A2 mixed definitions across cities (FUA extent for the Colombian cities, city extent for Accra, full lookup extent for the South African cities). The package now computes Table A2 under one uniform rule in the new script `part1_04c.city_table.R` (sourced by the master after `part1_03.config.R` and before `part1_05`): for each city, population, area, and subplace counts are computed from the final analysis dataset itself (`sum(population)`, `sum(area_km2)`, `nrow()`), and the FUA column is reported as N.A. for all cities, `part1_05` renders missing FUA values as "N.A." natively, so the five-column table layout is unchanged and no other script was modified. The script also writes the raw values to `outputs/tables/table_a2_values.csv` so the report table can be updated directly. The report's Table A2 is edited to match, with a footnote defining the new basis. The South African cities consequently show lower values than previously published (the analysis dataset filters to urban subplaces present in the OD matrix), and Metro Manila shows 1,877 units against the previously published 1,866.

11.2 South Africa paths

`private/zaf/Data_SAF_JAI/` is the canonical location for all eight South Africa inputs, and the code reads them from there (the original README's reference to `private/zaf/` directly was imprecise). Copies of the three subplace lookups may also exist under `South_Africa_Data_Prep/` as harmless leftovers of the prep workflow.

12. Data Sources (Data Availability Statement)

For every dataset, this section records the citation and source; a direct URL where one exists, or compilation steps where it does not; whether the data originate with a national statistics office (NSO), another agency, or the World Bank team; the access date; and a Development Data Hub (DDH) archiving flag for World Bank-originated data. Script read-locations are stated so provenance can be checked against the code.

12.1 Confidential raw data, read directly by the Part 1 pipeline

File	Source, contact, access date
`gha/econ_census_grid.gpkg`	Ghana Statistical Service, Economic Census 2023 at grid level. Shared confidentially by Anthony Krakah (anthony.krakah@statsghana.gov.gh). Accessed 10/02/2025. NSO.
`gha/WB_data_gridded_updated_20240213.csv`	Population census at grid level. World Bank Poverty GP, Laura Rodriguez Takeuchi (lrodrigueztak@worldbank.org); requestable from Statistics Ghana (census2021.statsghana.gov.gh, 2021phc@statsghana.gov.gh). Accessed 10/02/2025. WB-originated, flag for DDH.

File	Source, contact, access date
col/bogota/MA_GEO_JUL_25/MA_GEO_JUL_25.csv	Formal employment registry (Mercantiles) with coordinates. District Planning Secretariat of Bogotá, Andres Leonardo Acosta Hernandez (alacosta@sdp.gov.co). Accessed 10/12/2025.
col/SHP_Manzanas_CNUE/Manzanas_CNUE_Marco_Pos_DEST.gdb	Establishment counts (CNUE) with sector classification; same source. Accessed 09/17/2025. Used for both Bogotá and Cali. Layer Manzanas_CNUE_Marco_DEST_290722; 515,994 features; MAGNA-SIRGAS; key COD_DANE.
col/cali/base_prinicipal_corregida.xlsx	Cali business registry with sector and employment. Mayor's Office of Cali, Andres Prado Caicedo (andres.prado@cali.gov.co). Accessed 09/17/2025. 162,130 rows; join key cod_barrio (zero-padded character).

12.2 South Africa prep raw inputs (South_Africa_Data_Prep/, own README)

File	Source, access date
zaf/hex7_FTE_WageBand_Industry5d_Monthly.csv	SEAD-SA (Spatial Economic Activity Data, South Africa), derived from restricted SARS–National Treasury administrative tax data; the hex7 aggregate is the shippable form. Accessed 12/2025. Shared with the reproducibility unit.
zaf/asset_index.csv	Asset index (poverty proxy) at subplace level, derived from Stats SA Census 2011 subplace tables (info@statssa.gov.za; superweb.statssa.gov.za after registration). Accessed 07/2023.
zaf/subplace_population_ghsl100m.csv	Zonal summary of GHS-POP R2023A (100 m) to subplace polygons. DOI: 10.2905/2FF68A52-5B5B-4A22-8F40-C41DA8332CFE. Accessed 07/2023. Public source.
zaf/SP_SA_2011_clean.gpkg	Stats SA SP_SA_2011 subplace boundaries, cleaned; all 22,108 geometries validated. Accessed 07/2023. NSO.
zaf/UberH3_7.shp (+ .shx, .dbf, .cpg)	H3 resolution-7 hex grid, SEAD-SA portal. Accessed 07/2025. Ships without .prj; EPSG:4326 set explicitly by the prep script.

What each input supplies, and how the toolkit is built. The Spatial Tax file supplies employment (`fte_wage_0_2019`); the asset index file supplies the poverty proxy (`asset_index`); the GHS-POP zonal summary supplies `population`; and the subplace boundaries supply `SP_CODE`, district names, and geometry. Two prep scripts, run in order by `run_prep.R`, produce the toolkit: `01_make_weights.R` assigns each subplace to one H3 resolution-7 cell by centroid containment and weights it by its share of subplace area within the cell (areas computed in Africa Albers Equal Area); `02_make_toolkit.R` allocates 2019 hex-level employment to subplaces using those area weights and attaches the asset index and population. Area-only weighting prevents leakage of predictor features (building footprints, population density) into the employment outcome. The toolkit exists in two forms: `data_saf_global_toolkit.csv` in `Data_SAF_JAI/` is the frozen input read by `part1_02b`, and `data_saf_global_toolkit.rds` in the Part 2 Outcomes folder is the frozen input source consumed by `outcomes_southafrica.R`.

The files actually read by `part1_02b` sit in `private/zaf/Data_SAF_JAI/` (Section 11.2): the toolkit, the three subplace lookups (derived from the cleaned boundaries by district-municipality filter), and the three OSRM car OD matrices built by `od_matrix_southafrica.R`. A variable dictionary ships alongside them (`data_dictionary.txt`).

12.3 Philippines prep raw inputs (Philippines_Data_Prep/, own README)

All obtained under confidential terms from the Philippine Statistics Authority (PSA), National Censuses Service, contact Joseph P. Cajita (ncs.staff@psa.gov.ph), accessed 11/2025. A public-use file of the CPH 2020 is available through the PSA microdata portal (psada.psa.gov.ph) under registered access; the records used here were provided directly by PSA and are confidential.

File	Contents
`phl/{ncr,r3,r4a}_cph2020_form2.dta`	CPH 2020 Form 2 individual records for NCR, Central Luzon, and Calabarzon: population, working-age population, household education proxy.
`phl/cpbi2018_bgy_sector_wgt.dta`	CPBI 2018 weighted employment by barangay and 5-digit industry. Manila City barangays without establishment matches imputed by equal density within the sub-municipality.
`phl/phl_admbnda_adm4_psa_namria_20231106.shp` (+ sidecars)	PSA–NAMRIA ADM4 barangay boundaries, filtered to NCR and four adjacent provinces (Bulacan, Rizal, Cavite, Laguna).
`phl/ncr_bgy_match.xlsx`	Crosswalk reconciling CPBI legacy codes (prefix 1339) to PSGC 2023 codes (prefix 13806).
`phl/PSGC-3Q-2023-Publication-Datafile.xlsx`	Philippine Standard Geographic Code (3Q 2023), used to standardize codes across inputs.

What each input supplies. The three CPH 2020 Form 2 regional files (NCR, Region III, Central Luzon, Region IV-A, Calabarzon) supply barangay population (`pop_total`), the working-age population (`pop_25_65`), and the household education deprivation proxy (`deprivation_educ`, derived from census variable p17). The CPBI 2018 file supplies total employment and the four sector groups (`emp_sec_1`–`emp_sec_4`), with employment for Manila City barangays that do not match the establishment file imputed by equal density within the relevant sub-municipality (`employment_observed` preserves the raw pre-imputation join). The ADM4 boundaries supply the geometry, the `psgc10_shp`/`psgc9_shp` join keys, the administrative name fields, and `AREA_SQKM`; the crosswalk reconciles CPBI legacy codes (prefix 1339) to PSGC 2023 codes (prefix 13806); and the PSGC datafile standardizes codes across all inputs.

The Part 2 origins build. A single script, `build_ph_ncr_origins_fua.R`, takes these five inputs and produces `Outcomes/ph_ncr_origins_fua_10km.rds`, the file read by the frozen `outcomes_ph_ncr.R` prep step: it constructs the barangay-level population, deprivation, and employment measures, joins them to the polygons, identifies employment-center seeds, and builds the 10 km functional urban area around the NCR. Distances in this build are Euclidean throughout, the OSRM routing path is not used, so no routing server is required to reproduce this file.

The files read by `part1_02b` are the frozen prep outputs in `private/phl/JAI_Harmonized_PH_4_28_26/`, only the dated folder; the undated `JAI_Harmonized_PH/` folder is superseded and may be deleted. For reference, the harmonized barangay object `ph_jai_barangays_sf.rds` carries the columns the pipeline consumes: `barangay_id` (PSGC9 key, matching the row and column names of the OD matrix), `population_25_65`, `education_deprivation_index`, `employment_total`, the four sector columns (`emp_sec_1_production_infra`, `emp_sec_2_high_skill_private`, `emp_sec_3_public_social_services`, `emp_sec_4_retail_lower_skill`), and the barangay geometry.

12.4 Public raw data (Part 1)

File	Source, URL, access date
public/World Bank Official Boundaries - Admin 2.gpkg	World Bank official administrative boundaries. Sits at the ROOT of public/, not under gha/. Used to build accra_admin.rds.
gha/AMA/ama.shp (+ sidecars)	Accra Metropolitan Area boundary. World Bank Transport team, Fatima Arroyo Arroyo (farroyoarroyo@worldbank.org). No public URL. Accessed 04/08/2026. WB-originated, flag for DDH.
gha/accra.gtfs.zip	GTFS feed, Accra tro-tro network. TUMI Data Hub: hub.tumidata.org/dataset/trufi_gtfs_accra_accra. Accessed 04/12/2026. (The original README's filename accra.zip was incorrect.) Needed only if the transit cache is absent; the unzipped gha/accra.gtfs/stops.txt is needed on every run (Figure 6a).
gha/ghana-260412.osm.pbf	OpenStreetMap extract for Ghana, Geofabrik: download.geofabrik.de/africa/ghana.html. Accessed 04/12/2026. Needed only if the transit/walk caches are absent; OSM changes over time.
col/MGN2024_00_COLOMBIA/MGN/MGN_URB_SECTOR.shp	DANE Marco Geoestadístico Nacional: dane.gov.co/files/geoportal-provisional/, search "MGN2018 Integrado con CNPV2018, nivel de Sector Urbano" (12,169 features). Accessed 10/12/2025. NSO.
col/MGN2024_00_COLOMBIA/MGN/MGN_RUR_SECTOR.shp	As above, "nivel de Sector Rural" (8,163 features). Accessed 10/12/2025.
col/SHP_MGN2018_INTGRD_MANZ/MGN_ANM_MANZANA.shp	DANE, "MGN2018 Integrado con CNPV2018, nivel de Manzana Censal", CNPV2018 population joined to block geometry; roughly 2 GB extracted. Same portal. Accessed 10/12/2025. NSO.
col/VULNRB_IPMxMZ/VULNRB_IPMxMZ.shp	DANE, "Nivel de vulnerabilidad e IPM por manzana" (MPI at manzana level). Same portal. Accessed 10/12/2025. Listed here by provenance (public NSO data); stored under private/ in the package folder layout.
col/pm.xlsx	Municipal-level MPI, DANE (dane.gov.co, "medida de pobreza multidimensional de fuente censal"), prepared by WB consultant Julieth Andrea Parra Hincapié (julieth.parrah@gmail.com). Accessed 10/12/2025. WB-originated, flag for DDH.
col/mc_barrios/mc_barrios.shp	Barrio boundaries for Cali, Datos Abiertos Cali: datos.cali.gov.co/dataset/servicio-wms-barrios-de-cali ("Capa geográfica" → "Explorar" → "Ir al recurso"). Accessed 10/12/2025. id_barrio renamed to cod_barrio in the code.

Cali and Bogotá draw their poverty inputs from the same two files (the manzana-level MPI shapefile and the municipal pm.xlsx); there is no Cali-specific poverty source. These feed poverty_rate and, via weighted ntiles, the poverty quintile and decile columns.

12.5 Part 2 remote-sensed and downloaded sources

These openly accessible global datasets feed the Part 2 preparation workflows (Section 10). They are not redistributed raw; the processed per-city Features derived from them are the public data files of this package (CC-BY-4.0).

Dataset	Source, URL, access date
VIDA Google–Microsoft–OSM Open Buildings (per-country GeoParquet)	VIDA, on Source Cooperative; accessed programmatically via DuckDB httpfs in buildings_download.R (data.source.coop/vida/google-microsoft-osm-open-buildings/...). Accessed 01/2026. Basis for morphology features.

Dataset	Source, URL, access date
GHSL Building Volume, GHS-BUILT-V R2023A (100 m, 2020, Mollweide)	EC Joint Research Centre: human-settlement.emergency.copernicus.eu/ghs_buV2023.php. Accessed 01/2026. Read by raster_features.R.
GHSL Building Height, GHS-BUILT-H R2023A (ANBH, 2018, ~90 m, WGS84)	EC Joint Research Centre: human-settlement.emergency.copernicus.eu/ghs_buH2023.php. Accessed 01/2026. Verticality feature and gravity-mass input.
Nighttime Lights, VIIRS annual VNL V2 composite (2019, ~500 m)	Earth Observation Group, Payne Institute, Colorado School of Mines: eogdata.mines.edu/products/vnl/. Accessed 07/2023. NTL feature and its 2 km spatial lag.
Country OSM PBF extracts (ZAF, COL, GHA, PHL, VNM)	Geofabrik: download.geofabrik.de. Accessed 01/2026. Used by roads_extract.R and the OSRM OD scripts; roads are display/cost-benefit overlays only, excluded from prediction features.

12.6 Part 2 pipeline data folders (predictions/Data/)

Part 2 city tokens are exact and differ from Part 1 naming: Johannesburg, CapeTown, Durban, Bogota, Cali, Accra, PH_NCR, plus Hanoi for Features only.

Folder	Read by pipeline?	Contents and status
`Features/{city}_features.rds`	Yes	The nine remote-sensed predictors, the only public data files (CC-BY-4.0). Read by the labels script, all four engines, diagnostics, and prediction scripts.
`Outcomes/{city}_outcomes.csv`	Yes	poverty_rate, employment, population. CONFIDENTIAL (census/welfare-derived). Read only by part2_01_labels. Built by the outcomes prep scripts.
`OD_Matrices/{city}_od_wide.rds`	Yes	OD matrices read by part2_01_labels.
`Roads/{city}_roads.gpkg`	Optional	Label-map underlay only; NULL if absent.
`Rasters/, Clean_Boundaries/, Raw_Boundaries/`	No	Prep-only inputs. Not read by any pipeline script.
`Hanoi/ (8 subfolders)`	No	Intermediates of the Hanoi feature prep; only Features/Hanoi_features.rds crosses into the pipeline.

12.7 Datasets requiring DDH archiving (WB-originated)

Three datasets originate with the World Bank team and are flagged for Development Data Hub archiving: (1) the Ghana gridded population file `WB_data_gridded_updated_20240213.csv` (Rodriguez Takeuchi, Poverty GP); (2) the Accra AMA boundary `ama.shp` (Arroyo Arroyo, Transport); and (3) the Bogotá municipal MPI `pm.xlsx` (WB consultant Parra Hincapié).

12.8 Complete input-file inventory and classification

At the verification team's request, every file present in the data folder that is not produced by running the main pipeline has been classified. The full file-by-file inventory, with source, URL or access instructions, access date, citation, and license for every entry, is provided in the accompanying spreadsheet (JAI_input_classification.xlsx); this section defines the categories and summarizes the counts.

Category	Definition	Count
Real input	Read directly by the main pipeline from an external source (e.g., census microdata, DANE geographies, GTFS, official boundaries).	51
Frozen input	Pre-computed by a prep workflow, shipped as-is, and read by the pipeline (e.g., the SA toolkit and OD matrices, the harmonized PHL files, the Part 2 Features, Outcomes, and OD matrices).	38
Frozen input source	External data read only by a frozen prep workflow to produce a frozen input (e.g., the SEAD-SA tax file, the CPH census files, the GHSL rasters).	35
Cache file	Written by the pipeline on first run if absent; shipped so no routing engine is needed (the five Part 1 travel-time matrices).	5
Prep intermediate / by-product	Produced inside a frozen prep chain and consumed only by later prep steps, or a by-product not read at all. Not needed for replication; may be deleted.	88
Documentation	Data dictionaries, questionnaires, product notes, prep-run manifests, and sub-READMEs.	24
Superseded / duplicate	Stray copies and superseded working files (e.g., the undated JAI_Harmonized_PH folder, spare GTFS archives). May be deleted.	14
Not read by any package code	Present in the folder but consumed by nothing (e.g., additional MGN and AMA layers bundled with the same source downloads, the CPH Form 5 files). May be removed.	48

Two notes on the inventory. Sidecar files (.shx, .dbf, .prj, .sbn, .sbx, .cpg) carry the classification of their parent shapefile. And one pipeline input does not appear in the file listing, presumably because it is a .gdb container rather than a flat file: `private/col/SHP_Manzanas_CNUE/Manzanas_CNUE_Marco_Pos_DEST.gdb` (CNUE establishment counts, a Real input, read by `part1_02a` for both Bogotá and Cali; documented in Section 12.1).

13. License

The code is licensed under the MIT License together with the World Bank IGO Rider. The Rider is purely procedural: it reserves the privileges and immunities enjoyed by the World Bank without adding restrictions to the MIT permissions. The processed remote-sensed feature files, the only data files in the public package, are released under CC-BY-4.0. All other data are governed by the access conditions in Section 12 and are not redistributed.