Codebook

We use 26 different sources that we describe in very much detail in the README file. Since we also label all relevant variables of these data in Stata, we provide no additional codebook in cases where we republish the raw data as part of the replication directory. The README file documents the respective path for each raw dataset. Additionally, small README files in the respective folders of our replication directory provide additional guidance for those interested in more details.

Naturally, we also describe in detail the sources for which we cannot publish the raw data in the README file. This is true for five sources. In three cases, we cannot provide the raw form because of licensing issues and in two cases because of size restrictions of the replication directory. In all such cases, we provide transformations of the raw data in that we aggregate them to our city grids. In this codebook, we provide additional background on the raw files we obtained for these 5 datasets. We describe the names of the relevant variables from the datasets. Replicating some of the description of the README file, this codebook also briefly indicates which files in the replication directory were generated from these raw data and how.

NETS data

We obtained the following files in the ZIP archive from the data provider after purchasing the data:

1. "NETS2022_Emp_SIC/NETS2022_Emp.txt",
2. "NETS2022_Emp_SIC/NETS2022_NAICS22.txt",
3. "NETS2022_Misc/NETS2022_Misc.txt",
4. "NETS2022_Move/NETS2022_Move.txt" .

The relevant variables that we use in our paper are (original names in bold):

- **Dunsnumber**: the plant ID from Dun&Bradstreet
- employment (from NETS2022_Emp.txt) : **EmpXX** (where XX=90-22 is the year); this is employment of the establishment in year XX
- industry codes (from NETS2022_NAICS22.txt) : **NAICSXX** (where XX=90-22 is the year); these are 6-digit industry codes
- Latitude and Longitude (from NETS2022_Misc.txt and reconstructed using NETS2022_Move.txt): **Latitude**, **Longitude**; contains lat and lon information; they need to be reconstructed by year using the NETS2022_Move.txt move history file (see the code in extract_NETS).

Processing these input variables using the C++ program provided and documented in "_Work/C++/extract_NETS" will produce the grid-level aggregated outputs shared in the text files in the "_Data/US METRO DATA/CLUSTER OUTPUT" folder of our replication directory. *The readme file provides an exact description on how we process these data into both clusters and cell-level employment*.

Global Multi-resolution Terrain Elevation Data 2010

We do not provide the raw file, because the data are too large. At the same time, they are securely stored by the US Geological Survey (see README for all access details). The raster file associated with this dataset simply provides the mean elevation. The source and details on the processing are described in the README file.

The mean elevation is part of the developable definition. If slope>20% indicates non-developability. The README file provides the corresponding details. The results of the developable classification are saved in "_Data/125GLOBALCITIES/CLUSTERING/cells".

ESA World Cover 10m 2020

The ESA world cover raster is not included because of its size. The below table is a reprint from the Product User Manual (p. 13; licensed as CCBY 4.0). It describes the meaning of the raster codes (MAP code). The Manual can be accessed at the following URL: https://worldcover2020.esa.int/data/docs/WorldCover_PUM_V1.1.pdf (last accessed Sep 26, 2025).

Map code	Land Cover Class	LCCS code	Definition	Color code (RGB)
10	Tree cover	A12A3 // A11A1 A24A3C1(C2)-R1(R2)	This class includes any geographic area dominated by trees with a cover of 10% or more. Other land cover classes (shrubs and/or herbs in the understorey, built-up, permanent water bodies, ...) can be present below the canopy, even with a density higher than trees. Areas planted with trees for afforestation purposes and plantations (e.g. oil palm, olive trees) are included in this class. This class also includes tree covered areas seasonally or permanently flooded with fresh water except for mangroves.	0,100,0
20	Shrubland	A12A4 // A11A2	This class includes any geographic area dominated by natural shrubs having a cover of 10% or more. Shrubs are defined as woody perennial plants with persistent and woody stems and without any defined main stem being less than 5 m tall. Trees can be present in scattered form if their cover is less than 10%. Herbaceous plants can also be present at any density. The shrub foliage can be either evergreen or deciduous.	255, 187, 34
30	Grassland	A12A2	This class includes any geographic area dominated by natural herbaceous plants (Plants without persistent stem or shoots above ground and lacking definite firm structure): (grasslands, prairies, steppes, savannahs, pastures) with a cover of 10% or more, irrespective of different human and/or animal activities, such as: grazing, selective fire management etc. Woody plants (trees and/or shrubs) can be present assuming their cover is less than 10%. It may also contain uncultivated cropland areas (without harvest/ bare soil period) in the reference year	255, 255, 76
40	Cropland	A11A3(A4)(A5) // A23	Land covered with annual cropland that is sowed/planted and harvestable at least once within the 12 months after the sowing/planting date. The annual cropland produces an herbaceous cover and is sometimes combined with some tree or woody vegetation. Note that perennial woody crops will be classified as the appropriate tree cover or shrub land cover type. Greenhouses are considered as built-up.	240, 150, 255
50	Built-up	B15A1	Land covered by buildings, roads and other man-made structures such as railroads. Buildings include both residential and industrial building. Urban green (parks, sport facilities) is not included in this class. Waste dump deposits and extraction sites are considered as bare.	250, 0, 0
60	Bare / sparse vegetation	B16A1(A2) // B15A2	Lands with exposed soil, sand, or rocks and never has more than 10 % vegetated cover during any time of the year	180, 180, 180
70	Snow and Ice	B28A2(A3)	This class includes any geographic area covered by snow or glaciers persistently	240, 240, 240
80	Permanent water bodies	B28A1(B1) // B27A1(B1)	This class includes any geographic area covered for most of the year (more than 9 months) by water bodies: lakes, reservoirs, and rivers. Can be either fresh or salt-water bodies. In some cases the water can be frozen for part of the year (less than 9 months).	0, 100, 200
90	Herbaceous wetland	A24A2	Land dominated by natural herbaceous vegetation (cover of 10% or more) that is permanently or regularly flooded by fresh, brackish or salt water. It excludes unvegetated sediment (see 60), swamp forests (classified as tree cover) and mangroves see 95)	0, 150, 160
95	Mangroves	A24A3C5-R3	Taxonomically diverse, salt-tolerant tree and other plant species which thrive in intertidal zones of sheltered tropical shores, "overwash" islands, and estuaries.	0, 207, 117
100	Moss and lichen	A12A7	Land covered with lichens and/or mosses. Lichens are composite organisms formed from the symbiotic association of fungi and algae. Mosses contain photo-autotrophic land plants without true leaves, stems, roots but with leaf-and stemlike organs.	250, 230, 160

We extract the value to the central points of our grid cells using ArcGIS pro. Permanent water bodies and glaciers (70 and 80) are classified as non-developable. In case we had classified a cell as non-developable, but it is built-up (50), we reclassify it as developable. See README file provides more details. The grids containing the developable information are stored in “Data/US METRO DATA/GIS Data/US METRO GRIDS” (one file per city).

Commercial office stock data - Real estate investment trust (REITS)

The original data was provided by SNL research – the original provider was acquired by McGraw Hill, which is now S&P (see README for details). The file contains a large number of variables (including on the institutional investor, appraisal value etc.). The only information we use from this file are the coordinates:

- Latitude: **lat**
- Longitude: **Lon**

After a match with our grids, a simple count of establishments within grid cells results in the "grid_STarSNL.dta" data set shared in the "_Data/125 GLOBAL CITIES/STARBUCKS_SNL" folder of our replication directory.

Tweets data

The original data (business-related tweets) was acquired from GNIP - a company that no longer exists (see README for details). The file contained a large number of variables of which we only used those identifying the coordinates of the tweet:

- Latitude: **lat**
- Longitude: **Lon**

After a match with our grids, a simple count of establishments within grid cells results in the "DATA.dta" data set shared in the "_Data/125 GLOBAL CITIES/TWITTER" folder of our replication directory.

Geotaggers data

Erica Fisher shared the geo data with us. To preserve her artwork, we have matched the original data to grids. The raw data we had obtained from Erica Fisher are photo uploads that have variables for latitude and longitude. The relevant information from this dataset are:

- Latitude: **lat**
- Longitude: **Lon**

After a match with our grids, a simple count of establishments within grid cells results in the "PHOTOCOUNT_250M.dta" data set shared in the "_Data/125 GLOBAL CITIES/ PHOTOS" folder of our replication directory.