Estimating Functional Urban Areas without Commuting Data

Pedro Afonso
PARIS LODRON UNIVERSITY OF SALZBURG

January 15, 2026

Cities, like dreams, are made of desires and fears...
-Italo Calvino

Abstract

The number of people living in the commuting zones of large cities currently accounts for almost a fifth of the global population, however these populations remain poorly represented in global urban datasets due to the widespread scarcity of commuting data. In this project, I replicated and adapted the machine-learning approach of Moreno-Monroy et al. (2021) to estimate Functional Urban Areas (FUAs) worldwide using only globally available geospatial indicators.

I trained a logistic regression model on EU–OECD FUAs using distance to the city center, accessibility, Urban Center population, population density and sub-national Human Development Index (sHDI). The final model achieved an accuracy of 0.71 and a sensitivity of 0.69 on held-out FUAs, producing plausible commuting zones across a wide range of cities.

Keywords: Cities; Metropolitan Areas; Functional Urban Areas; Suburbanization

1 Historical Context

Comparing cities is not a straightforward task, since the very definition city varies greatly from country to country, even in the same global region. For this reason, in 2012, the Organisation for Economic Cooperation and Development (OECD), in cooperation with the European Union (EU), created an harmonized definition for city and metropolitan area (OECD, 2012). The method can be summarized in the following way: cities are composed of the agglomeration of high-density 1km^2 cells, and, afterwards, the surrounding municipalities where at least 15% of the working population commutes to the city core are considered part of its commuting zone; together they form the metropolitan area of said city, in this context called the Functional Urban Area (FUA).

Several papers were written on the subject and this definition was widely use for comparison between cities of the aforementioned organizations. However, as late as 2019, the method was still fully dependent on commuting data (Dijkstra et al., 2019), which effectively excludes most of the world.

This was finally handled in 2021, when the OECD began implementing a machine-learning approach to estimate the FUA of any city in the world, without the need of commuting data (Moreno-Monroy et al., 2021). I believe this might be a great resource in the context of GeoHumanitarian Action, since Moreno-Monroy et al. (2021) estimate that 17% of the world population lives in the commuting zones of big cities, meaning they cannot, therefore, be ignored when studying subjects such as the propagation of epidemics, the access to health and education, and the overall propagation of any city crisis into the surrounding regions. For this reason, I tried to replicate the method described by Moreno-Monroy et al. (2021).

2 Method Overview

The method implemented by Moreno-Monroy et al. (2021) starts from the urban clusters of high-density cells described in Section 1, since this part doesn't require commuting data. They then estimate how likely it is for each cell in the globe to be part of a neighboring city's FUA using a logistic regression model trained with information on actual EU-OECD FUAs, and agglomerate cells above a certain probability threshold.

Only cells with a population of 300 inhabitants or more

are considered on this step, and the model is trained using the distance of the cell to the urban center, the size of the urban center, and the cell population. In addition to this, they calibrate the model by world region and using country-level GDP per capita to «help mitigate a possible bias in the country sample used for the training set».

At last, to avoid colossal urban centers (as would happen in places like Egypt and Bangladesh, due to their concentrated populations), they impose two additional rules:

- FUAs can only have one urban center of 500.000 or more inhabitants;
- urban centers with more than 20 million inhabitants and more than 2.500 km^2 are split if they have at least two hypercores, which are agglomerations of super-high-density cells containing at least 1 million people.

These two rules are the only thing I have yet to implement, because the recurring nature of the first rule and the recalculation of urban centers in the second rule would be extremely time-consuming. That being said, I will possibly do so in the future to fully complete the project.

3 Data Sources

A lot of different datasets were needed for this endeavor. The first, and possibly the most important, was the set of EU-OECD FUAs, which can be downloaded from the OECD website. As for the population data, I used the population density data from GHS-POP R2023A (Schiavina et al., 2023a) and the urban center boundaries from the GHS Urban Centre Database, derived from GHS-SMOD R2023A (Schiavina et al., 2023b), both within the Global Human Settlement Layer framework (Pesaresi et al., 2024).

Another dataset used by Moreno-Monroy et al. (2021) was the global travel impedance grid (Weiss et al., 2018), which also consists of a $1km^2$ global grid, with the accessibility to urban centers. They then used «information on roads, railroads, water bodies and movement over land» to calculate the actual travel time from each cell to the urban centers. I tried to accomplish the same and was not able to replicate it, so, instead, I used the global travel impedance grid in conjunction with the "straight-line distance" calculated through conventional methods.

As already mentioned, they also used country-level GDP per capita, but I decided to use the Human Development Index instead, for a couple of reasons. First, I wanted to use sub-country-level index that would provide a wider interval of values and also allow the model to distinguish cities in poor regions from cities in richer ones (in my opinion, this is especially important in developing countries, where the capital city is often much richer than the rest of the country); I couldn't find a sub-country-level GDP per capita dataset, but I did find one for Sub-National HDI (SHDI) (Smits and Permanyer, 2019). Secondly, I don't think that using the SHDI introduces any errors, because, despite their conceptual differences, it's been shown that HDI has a very high correlation with GDP per capita (Sušnik and van der Zaag, 2017).

Furthermore, I used a high-resolution shapefile of country borders, which can be downloaded from the World Bank Group website; and also the OECD FUA predictions (Schiavina et al., 2019), made available via the Global Human Settlement Layer framework, for comparison in the end.

All of these datasets used the same Coordinate Reference System, `EPSG:4326`, except the population density data and the urban center boundaries, which use a custom projection called `World_Mollweide`, that has a cartesian coordinate system, in contrast with the elliptical system of `EPSG:4326`.

4 Implementation

4.1 Preparing the data

The first step was to load all the vector datasets and convert them to `World_Mollweide`. Then, since the dataset of the urban center boundaries didn't have the ISO Country Codes, I had to manually rename all the countries with mismatching names across datasets.

With everything in accordance, I was finally ready to start combining the data. I began by calculating the intersection between the EU-OECD FUAs and the urban centers' dataset, and then assessed what was the largest urban center in each FUA. Said urban center would then be considered the core of the FUA, and so I added two new columns to the EU-OECD FUAs dataset called `biggest_city_id` and `biggest_city_pop`, dropping FUAs where no urban center was found.

Afterwards, I did the same for the HDI and created a new column called `biggest_city_shdi`. In both cases, I had to be very careful with urban centers intersecting multiple FUAs or multiple HDI regions. For example, before I fixed this problem, I had Ceuta (Spain) accidentally using the HDI of Northwest Morocco. With this finished, I was left with 1075 FUAs out of the original 1311.

I then created histograms to analyze the distribution of values for `biggest_city_pop` and `biggest_city_shdi`, and noticed there was a huge imbalance in the population histogram, with 90% of urban centers having less than 1

million inhabitants and only 1% having more than 10 million. Considering that it is normal to leave out 20% of the dataset for testing, there was a big chance it would exclude all megacities, worsening the model predictions later on. For this reason, I decided to follow the advice of Geron (2019) and create 10 "population categories" with intervals of 500.000 people (with the exception of the highest one, which encompassed all the remaining urban centers), that would then be used to proportionally select my random samples.

4.2 Feature engineering

Moreno-Monroy et al. (2021) used the logarithm (with varying powers) of each feature for the training of the model. I initially did the same, but wasn't achieving good results, particularly with the coefficients associated with each feature, which were often the opposite sign of what one would expect: accessibility (which is defined to be proportional to travel time) was getting a positive coefficient, while population density was getting a negative one.

I thus began experimenting: I tried selected use of the logarithm, complete exclusion of logarithms, and even alternatives such as square roots and other decimal powers. While, the coefficient of the accessibility was quite easy to fix, nothing was yielding a positive coefficient for the population density. After this, I even ran the script without population density entirely a few times, but the accuracy and sensitivity were even lower. As a last resort, I decided to attempt a scaled population density instead, i.e. using the population density relative to the urban center's own density... And it worked! Using the scaled population density represented a paradigm shift: the model was no longer evaluating if a cell was very dense or not, but wether it was dense in the context of its surroundings.

Admittedly, I stopped experimenting after this, which means this might not be the most optimized model possible. Nonetheless, I settled on using the transformations shown on Table 1.

Feature	Transformation Applied
Urban Center Population	linear
Population Density	$\frac{\rho}{\rho_{\mathrm{UC}}}$
Accessibility	linear
Distance	$\log(1+d)$
HDI	linear

Table 1: Transformations applied to each feature during the logistic regression model training.

4.3 Feature Extraction and Training

The largest difference between my approach and the method described by Moreno-Monroy et al. (2021) was that I didn't use the entire world's population density grid at the same time, for a very simple reason: my computer was incapable of doing so. With this little setback, I had but one choice: go through the FUAs one by one and load a smaller windows for each of them instead. But how large should this window be? Should it be the same for all the FUAs?

For reference, I confirmed how big the largest FUA in my dataset was; it was Los Angeles (United States), with a FUA of nearly 100.000 km^2. At first, I decided to load a window of twice that much (to take into account the irregular shape of some FUAs), then I downgraded it to the 100.000 km^2, but the feature extraction process was still taking half a dozen hours and 80% of my cells were not part of any FUA, so my sample was extremely biased, and the accuracy assessment hard to rely on. Finally, I decided to load a custom window for each FUA. I experimented with different values and settled on a window 4 times the FUA's size, while also enforcing a ceiling of 100.000 km^2. The speed of the feature extraction process increased fivefold.

Another thing to consider was what the exact center of my window should be; the same center that I would then use to calculate the "straight-line" distance. At first, I picked the biggest urban center's centroid, however, for large coastal cities that grew inwards with time, the centroid ends up quite far from the actual center of the city, to where most people are actually commuting. To fix this, I implemented a small pipeline that would consult the city's coordinates using Nominatim, which uses OpenStreetMap data to find locations on Earth by name (and vice-versa), and would then verify if the given coordinates fell inside my urban center's boundaries. If so, it would take that as a valid point and use it as the center; if not, it would simply take the centroid. Around 1 for every 7 FUAS ended up using the centroid, which mostly happened on small urban centers where the coordinates fell just outside the border. An example of this can been seen in Figure 1.

The final issue I had to take into account for the training was the fact that some urban centers included two cities with distinct FUAs on the OECD-EU dataset. Such cases include Rotterdam-The Hague (in the Nethelands) and Wiesbaden-Mainz (in Germany). To fix this, I excluded any areas of the urban center outside the FUA, and recalculated the area and population.

With all this under consideration, my script went through the FUAs one by one (as mentioned before), loaded the core city (with its population and HDI), loaded an appropriately-sized window of the population density

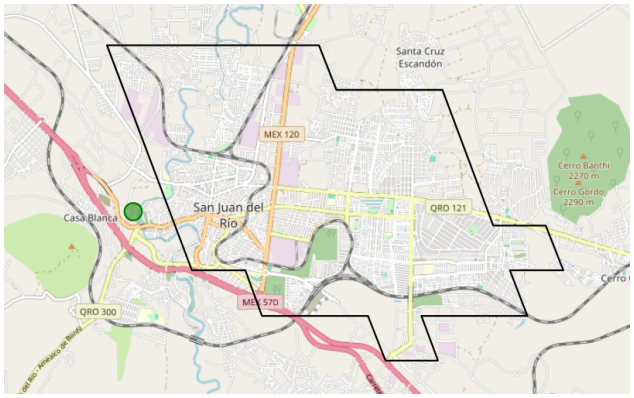

Figure 1: San Juan del Río, in Mexico, with the urban center outline in black, and the Nominatim center to the left, in green.

and accessibility rasters (with special care to make sure they were not misaligned), eliminated inaccessible cells, off-country cells and cells with a population density below 300 inh./km^2,[1] and then calculated the distance to each of the remaining ones.

Once the features of all the FUAS were extracted, the model could be trained.

4.4 Model Calibration

Similarly to Moreno-Monroy et al. (2021), I also calibrated the model once it was trained by finding the probability threshold that maximized accuracy with my testing set. The difference is that they chose a different threshold for each world region, whereas I chose a global one, hoping that the HDI feature used in the training would be enough to differentiate between regions.

My final accuracy was 0,71, while my final sensitivity (recall) was 0,69.

4.5 Final Adjustments

With the model calibrated, I could now predict which surrounding cells belonged to the FUA of any urban center. All that was left to do was to turn the resulting assortment of cells into an actual polygon. To do so, I started by aggregating the non-FUA cells using alpha-shapes (Edelsbrunner et al., 1983) with $\alpha = 1\backslash 2000$ (which defines a generalized disc of radius 2000m, i.e twice my grid resolution). Then, I recovered my raster grid and excluded all the FUA polygons within the non-FUA alpha-shapes, therefore eliminating any outliers. Finally, I aggregated the remaining FUA cells using alpha-shapes with $\alpha = 1\backslash 5000$, a value lower than before, to heavily decrease the chance of FUAs being multi-polygons, but which can be adjusted if one sees fit. The FUA was the union of the resulting shape and the Urban Center, to ensure that I included the Urban Center cells that, for some reason, had been excluded so far.

I also experimented with the option of disregarding country borders, which seems reasonable in contexts such as the Schengen Area, and achieved interesting results for cities such as Strasbourg and Salzburg.

4.6 Conclusion

I think I completed all the objectives I set out to do when I first began this project, i.e. I obtained a working model capable of predicting the Functional Urban Area of a city with relative accuracy (71%), which even surpasses the original one at times, for example by including Chantilly in the FUA of Paris (France).

Distance is, by far, the most important feature in my model, with the radius of influence depending on the HDI and the urban center's population (and increasing with both). The accessibility and population density appear to act mostly close to the edge, thus helping to define the border.

[1] I also experimented with excluding the cells inside the urban center during training, since those are already part of the FUA by definition, but this made the model perform worse and reverted the sign of the population density's coefficient as well.

References

Dijkstra, L., Poelman, H., and Veneri, P. (2019). *The EU-OECD definition of a functional urban area*. OECD.

Edelsbrunner, H., Kirkpatrick, D., and Seidel, R. (1983). On the shape of a set of points in the plane. *IEEE Transactions on Information Theory*, 29(4):551–559.

Geron, A. (2019). *Hands-on machine learning with scikit-learn, keras, and TensorFlow*. O'Reilly Media, Sebastopol, CA, 2 edition.

Moreno-Monroy, A. I., Schiavina, M., and Veneri, P. (2021). Metropolitan areas in the world. delineation and population trends. *Journal of Urban Economics*, 125:103242.

OECD (2012). *Redefining "Urban": A New Way to Measure Metropolitan Areas*.

Pesaresi, M., Schiavina, M., Politis, P., Freire, S., Krasnodębska, K., Uhl, J. H., Carioli, A., Corbane, C., Dijkstra, L., Florio, P., Friedrich, H. K., Gao, J., Leyk, S., Lu, L., Maffenini, L., Mari-Rivero, I., Melchiorri, M., Syrris, V., Van Den Hoek, J., and Kemper, T. (2024). Advances on the global human settlement

layer by joint assessment of earth observation and population survey data. *International Journal of Digital Earth*, 17(1).

Schiavina, M., Freire, S., and MacManus, K. (2023a). Ghs-pop r2023a - ghs population grid multitemporal (1975-2030).

Schiavina, M., Melchiorri, M., and Pesaresi, M. (2023b). Ghs-smod r2023a - ghs settlement layers, application of the degree of urbanisation methodology (stage i) to ghs-pop r2023a and ghs-built-s r2023a, multitemporal (1975-2030).

Schiavina, M., Moreno-Monroy, A., Maffenini, L., and Veneri, P. (2019). Ghs-fua r2019a - ghs functional urban areas, derived from ghs-ucdb r2019a, (2015), r2019a.

Smits, J. and Permanyer, I. (2019). The subnational human development database. *Scientific Data*, 6(1).

Sušnik, J. and van der Zaag, P. (2017). Correlation and causation between the un human development index and national and personal wealth and resource exploitation. *Economic Research-Ekonomska Istraživanja*, 30(1):1705–1723.

Weiss, D. J., Nelson, A., Gibson, H. S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., Dalrymple, U., Rozier, J., Lucas, T. C. D., Howes, R. E., Tusting, L. S., Kang, S. Y., Cameron, E., Bisanzio, D., Battle, K. E., Bhatt, S., and Gething, P. W. (2018). A global map of travel time to cities to assess inequalities in accessibility in 2015. *Nature*, 553(7688):333–336.