Functional Urban Areas – Application of the Concept on Traditional Longitudinal Data for Intra- and Inter-City Analysis

Wenxiu Du[1], Andrew Ding[2], Dorothee Beckendorff[1], and Mathias Lerch[1]

[1]Urban Demography Laboratory, EPFL
[2]University of Waterloo

October 31, 2023

1 Introduction

The use of satellite-based data has greatly improved our understanding of the world. Nowadays, researchers use satellite data to perform a wide range of tasks, from predicting weather conditions and planning city landscapes to determining human settlement distribution and built-up surfaces. In demography, urban planning, and city science, one of the most important satellite-based datasets to consider is the Global Human Settlement Layer (GHSL), which contains fine-grid longitudinal raster data on population distribution, degree of urbanisation, and built-up characteristics, as well as many other datasets, such as Functional Urban Area (FUA), which describe and characterise human presence on Earth. However, the image-based data does not show individual-level information, therefore with no qualitative insights on who lives in a given city, how the population grow etc, which are only possible through surveying or data collected by governments. Among individual-level data, census data, which are extensively used around the world for planning, are especially rich in information, as they contain individual-level data that are representative of the corresponding country. Due to privacy issues, census data are often anonymised and grouped into administrative regions, so that individuals cannot be identified. Administrative borders are frequently changed by the governments, making longitudinal research on a particular region challenging to conduct. On top of that, census data are country-specific in terms of defining the criteria for classifying an area as urban or rural and whether an area belongs to a city or metropolitan area. These factors make cross-national longitudinal studies challenging when performing urbanisation or city-related studies. Because the newly available satellite-derived data are rich in nature and enable analyses that were previously impossible to perform, it is empirical to harmonise the geographical divisions of each country through different observation periods and to develop novel approaches of integrating newly available satellite-derived data into research using traditional data. This research is an attempt to bridge this gap between satellite-derived GHSL data and census data for longitudinal urban studies and will use all countries to their finest geographical division available.

2 Terminology

City proper, urban aggregation, urban agglomeration, functional urban areas and metropolitan area are often used to describe cities, but the exact differences are somehow blurred. In this research, they will be given a clear definition to describe different types of definitions of cities. Hereafter, city proper refers to an administrative locality with fixed boundaries that have been recognized as "urban" by the corresponding government (United Nations Department of Economic and Social Affairs, 2006). Urban aggregation refers to a continuous built-up area that contains one city core or several cities with shared land use of industry, infrastructure and housing (Loibl et al., 2018). Urban agglomeration contains a city proper and its suburban areas adjacent to the city proper (United Nations Department of Economic and Social Affairs, 2006). Functional Urban Areas (FUAs) contain a high-density urban centre and its commuting zone, where at least 15% of the inhabitants work in the urban centre (OECD, 2012). The metropolitan area is interchangeable with FUAs. By these definitions, the city proper's administrative borders are usually constant over time and do not take into consideration the spatial extension of the city. Therefore, the city proper is not an accurate representation of the city limit. Urban

aggregation and urban agglomeration are distinct from one another, the former consists of several cities that could be considered as one big urban area, whereas the latter focuses primarily on one single city and its suburbs. FUAs also differ from urban aggregation and urban agglomeration, as FUAs could contain areas that are not related to the city proper by a continuous built-up area, but through commuting flows which determined the border of which a certain percentage of inhabitants commute to the city centre to work. FUAs are arguably the best at identifying the boundaries of cities from a people-centric standpoint since they pinpoint the area in which people are strongly connected to a particular city centre.

3 Data

The satellite-derived data in this study originates from Global Human Settlement Layers (GHSL), which is an open and free database for assessing human presence on the planet (European Commission, 2023). This analysis will make use of three GHSL datasets. The first dataset is the GHS Population spatial raster, which depicts the distribution and concentration of the population on Earth, up to a resolution of 100m grid. This dataset is generated by using the population counts of each administrative area from census data and built-up characteristics developed by GHSL to redistribute the total population of each area to grid-cells (European Commission, 2023). In this research, 1km grid is employed to keep consistency with the resolution of other datasets. The second dataset is the GHS Settlement Model Layer (SMOD), which uses population size, population and built-up area densities to determine the degree of urbanisation (DoU) of each 1km-by-1km grid cell(European Commission, 2023). In this DoU framework, each grid cell is classified into one of the eight categories: 1) Urban Centre; 2) Dense Urban; 3) Semi-Dense Urban; 4) Suburban; 5) Rural; 6) Low Density Rural; 7) Very Low Density Rural; 8) Water (European Commission, 2023). Figure 1 shows the DoU of Geneva and the surrounding areas, as a visual illustration of the dataset. This definition breaks the rural-urban dichotomy that has been adopted widely by governments and researchers. The DoU captures the urban-rural continuum that is more aligned with reality, as many regions lie within the transitional stage between urban and rural, and a continuous measurement between these two can represent urbanisation more accurately. On top of that, GHSL uses the same criteria on population size and density to classify DoU across the entire world, disregarding nation-specific definitions of what regions' national governments qualify as urban or rural, making it much easier and more consistent for internationally comparative studies (Eurostat, 2021).

The third dataset is GHS-FUA Functional Urban Areas. Moreno-Monroy et al. (2021) used urban centres from SMOD plus their commuting zones to determine the boundary of each FUA, by considering urban centres with at least 50,000 inhabitants or with a density greater than 1,500 individuals per square kilometre. The commuting zone information from 31 OECD countries was available to the authors directly through government registers or indirectly through population and employment registers or mobile phone data and was employed as the training set for predicting the extent of FUAs in countries without commuting zone information (Eurostat, 2021; Moreno-Monroy et al., 2021). Figure 2 shows the FUA area of Rio de Janeiro, Brazil. The graph shows that the FUA of Rio de Janeiro is much bigger than the municipality of Rio de Janeiro, which is located in the southeast of the shaded area. Smaller cities such as Magé, Itaguaí and Petrópolis are also included in the FUA. A closer look into these cities shows that the train line Guapimirim links Magé to Rio de Janeiro, and Petrópolis and Itaguaí have regular bus services to Rio, indicating strong connectivity between these regions to Rio. This definition is more advantageous than considering city proper or urban agglomeration, as it considers the border of cities based on people's connection to a certain city centre, and it minimises the under-estimation of the city population due to urban sprawl, a process in which inhabitants move out of the city centre into surrounding regions while maintaining strong ties with the city. Furthermore, the FUAs are globally comparable and do not depend on city-proper definitions because the data is based on GHSL, making it suitable for cross-national city-level analysis. This GHS-FUA will be used to define FUAs adapted to the administrative geographies of each country. Note that GHS-FUA is only available for the year 2015 for the time being, therefore the following analysis is based on GHSL data from 2015.

The fourth dataset is the Database of Global Administrative Areas (GADM), which includes the administrative division of all countries/regions on Earth (GADM, 2023). The GHS-FUA is matched onto administrative divisions, rather than used directly, since publicly available data is often aggregated by administrative divisions due to data protection laws and privacy issues, making it impossible to find out the exact location of each individual in the dataset. The GHS-FUA often cuts through administrative areas, making the direct use of GHS-FUA on administratively aggregated data impossible without transformation. Therefore, the FUA boundary adapted to

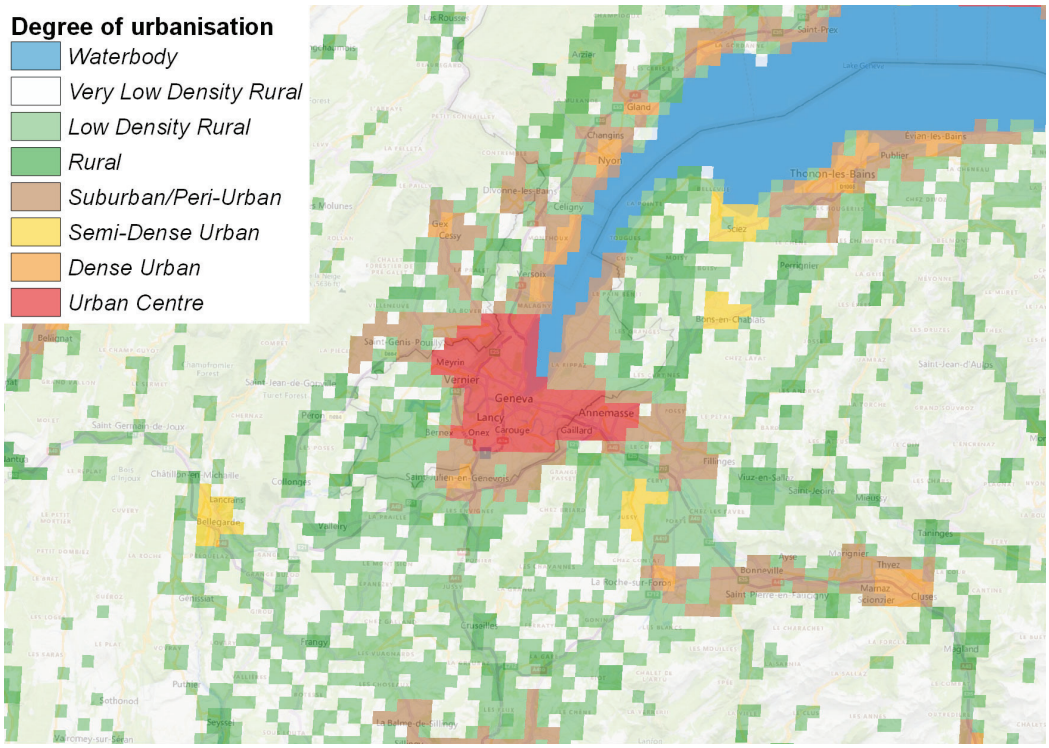

Figure 1: Degree of Urbanisation: Geneva and Surroundings

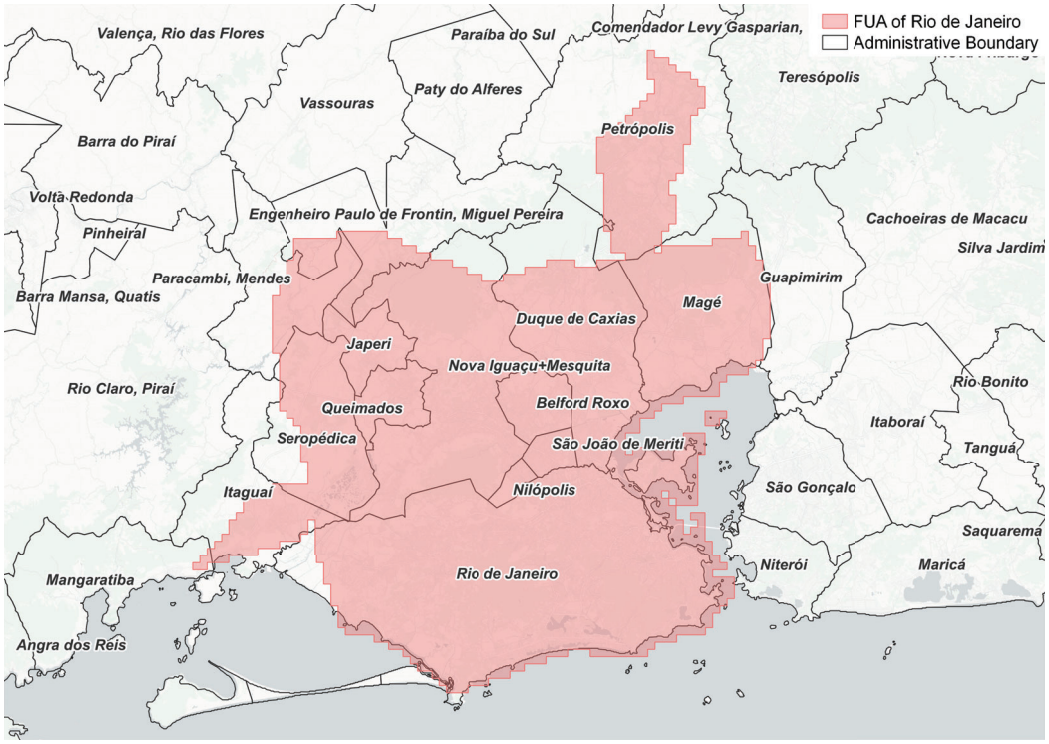

Figure 2: Functional Urban Area of Rio de Janeiro

administrative geographies needs to be determined to facilitate the integration of GHS-FUA and traditional data.

The fifth dataset is the collection of GIS boundary files from Integrated Public Use Microdata Samples (IPUMS), which contains both year-specific and spatially harmonised first, second and third level geography for a selected group of countries (Minnesota Population Center, 2019). This is the geographical base of the URBDEMO collection of IPUMS population and housing censuses between 1980 and 2017 for 40 countries in the Global South, maintained at the Urban Demography Lab of the Ecole Polytechnique Fédérale de Lausanne (EPFL). Therefore, the countries that are being considered in this communication are Burkina Faso, Benin, Bolivia, Brazil, Botswana, Chile, China, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, Egypt, Guatemala, Honduras, Indonesia, Iran, Kenya, Cambodia, Morocco, Mali, Mongolia, Malawi, Mexico, Malaysia, Mozambique, Nicaragua, Nepal, Panama, Peru, the Philippines, Paraguay, Sierra Leone, Senegal, Suriname, Thailand, Tanzania, Uganda, Uruguay and Venezuela. In this preliminary analysis, this set of data will be used to illustrate the results of the following algorithm, and GADM data will be integrated in the final communication.

4 Methodology

4.1 Geographical Harmonisation

Geographical harmonisation is an important factor to consider when analysing longitudinal data at the administrative level, as governments often change administrative boundaries for various reasons, making it difficult to conduct meaningful longitudinal studies. Some of the challenges caused by these boundary changes include not being able to find the same administrative area from data of one year to another, or observing a significant demographic shift in a region between two or more years caused only by boundary changes. Some datasets, such as that by IPUMS, have harmonised geographical units, but they are often produced using geographic definition over a long period of time, and cannot to adjusted to a subset of the years of interest. This is particularly limiting when researchers want to perform analyses focusing on small geographical scales, as the merging often creates large geographical units. It is therefore beneficial to create an algorithm that defines a consistent geographical delineation of administrative units over time, in order to focus the analysis on demographic dynamics within constant spatial units.

The following algorithm is a viable and repeatable method for harmonising the administrative geographies across different points in time. The algorithm is recursive, i.e. meaning that it is repeated until the condition is fulfilled, and it merges adjacent geographical units that have changed boundaries between two or multiple observation points until it reaches stable boundaries that are no longer affected by the boundary changes.

1. Compare all geographical units between year 1 and year 2. If a geographical unit from year 1 can be found in year 2 with the same boundary, then add this unit to the "harmonised" list for year 1, with an indication of which unit in year 2 it is being matched to. Otherwise, add the unit to the "to be harmonised" list for year 1.
2. Go through all the units in the "to be harmonised" list for year 1:
   (a) If one unit from year 1 has the same boundary as a combination of units in year 2, then add that one unit from year 1 to the list of "to be harmonised" for year 1, with a note of which units in year 2 it is being harmonised with. Check all units from year 1 before proceeding to the next step;
   (b) If the combined two or more units from year 1 are the same as one unit from year 2, then combine these units from year 1 and add them to the list "harmonised" for year 1, with a note to which unit in year 2 it is being matched. Check all units from year 1 before proceeding to the next step;
   (c) If the combined two or more units from year 1 are exactly the combined two or more units from year 2, then combine these units from year 1 and add them to the "harmonised" list for year 1, with a note to which unit in year 2 it is being matched.
   (d) Stop the algorithm when all the units in year 1 are on the harmonised list.
3. Repeat for all combinations of censuses until the result converges and the list "harmonised" does not change for all censuses and all combinations.

Figure 3 shows an example of geographical harmonisation using the GIS boundary files of the Dominican Republic for 1981, 2002 and 2010. The coloured lines correspond to the year-specific boundaries and the black

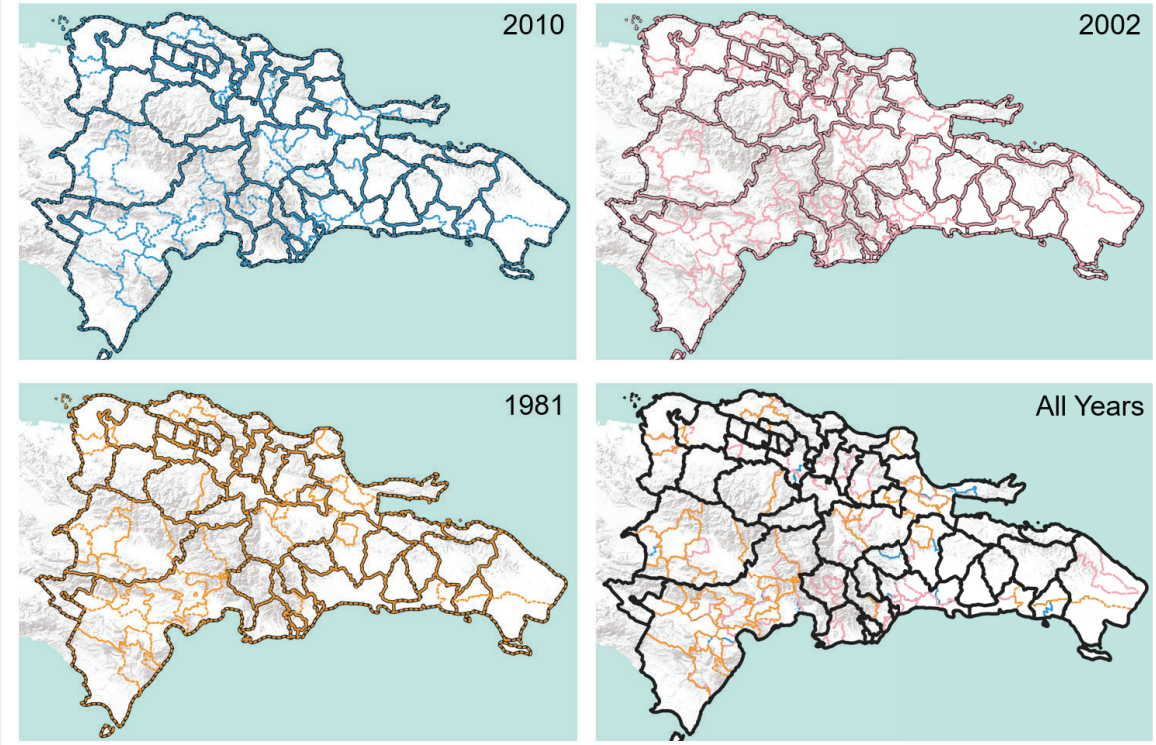

Figure 3: Geographical Harmonisation of GIS Boundary from the Dominican Republic across 1981, 2002, 2010 - Coloured line indicates year-specific boundary, black line indicates harmonised boundary

line corresponds to the harmonised boundaries using the algorithm mentioned above. The algorithm captures small changes in boundaries that may not be visible at the country level and provides satisfactory results overall.

4.2 GHS-FUA Matching Algorithm

As a next step, we assign each harmonised administrative area to the FUA within which it is situated. The matching takes into consideration population and inhabited area. For each administrative area, its inhabited area and the spatial distribution of the population within that area are determined using the GHS-POP raster of the year 2015 overlayed over the administrative geography shapefile. The inhabited area is determined by calculating the area in which the GHS-POP is not 0, i.e., the total area of 1km of grid cells that are inhabited.

Each administrative area polygon is then intersected with a GHS-FUA polygon to determine the population and the inhabited area in each of the intersections. For each intersection, three parameters are calculated to decide whether a given administrative area is part of a specific GHS-FUA:

1. The percentage of the administrative area's population that is located within a specific GHS-FUA;
2. The percentage of the administrative areas' inhabited lands that are located within a specific GHS-FUA;
3. The percentage of the number of individuals living in a specific FUA that also live in a specific administrative area (Number of individuals of an FUA/Number of individuals living at the intersection between that FUA and an administrative area).

Using these parameters, each administrative area is then iteratively assessed as belonging to a GHS-FUA or not. In order to distinguish the GHS-FUA defined by GHSL and those being matched using administrative areas, the former will be referred to as oFUA (original FUA) and the latter as mFUA (matched FUA) hereafter. The algorithm identifies the geographical units that have at least 50% of their population living in a particular oFUA and takes into account different criteria using a combination of the three parameters mentioned above to identify mFUAs that are closely matched to the oFUA and those that are not so well matched but still could be considered as matches to a particular oFUA.

There are three different levels of matching results:

1. Level 1 - High-quality Matches: In this level, the intersecting population and inhabited areas are considered between each oFUA administrative area pair. If parameter 2 is greater than 50%, and parameter 1 is greater than 50%, meaning that the majority of the population in an administrative area lives in an oFUA, and the majority of inhabited land of an administrative area is located within an oFUA, then the pair is given scores above 95 and assigned to the corresponding mFUA. This matching score corresponds to the purple area in Figure 4. Subsequently, for each mFUA, we further identify geographical units that have parameter 1 greater than 50% and parameter 2 smaller than 50%, meaning that the majority of the population in an administrative area lives in an FUA, then these administrative areas are also incorporated into the identified FUAs and given a score between 85 and 89. This corresponds to the green part in Figure 4.
2. Level 2 - Intermediate quality Matches: In this level, parameters 1 and 3 are utilised. For oFUAs that are not being matched in level 1, and if the intersection has parameter 1 greater than 50% and parameter 2 greater than 50%, meaning that the majority of the population in an administrative area lives in an FUA and the majority of the population of that FUA also lives in that administrative area, then it is given scores between 75 and 79 respectively. Two parameters were used to make sure the geographical unit is a good representation of a particular FUA, meaning that most people of that FUA live in that geographical unit and vice versa. This is not matched in level 1 often because level 1 matching requires at least one geographical unit that matches well both in terms of population and inhabited area. At this level, the inhabited area is not considered. This corresponds to the pink part in Figure 4.
3. Level 3 - Low-quality Matches: In this level, only parameter 1 is used. For each administrative area, the sum of its parameter 1 across all oFUAs not yet being matched is calculated. This corresponds to the total number of individuals in a particular geographical unit living in a not-yet-matched oFUA over the total population of that geographical unit. If this parameter is greater than 50%, then it is given scores between 65 and 69. If multiple FUAs are being identified for a particular administrative area, then the oFUAs are grouped and form a new mFUA. If the newly grouped oFUA in this level also intersects another administrative area, and if this area has more than 50% of its population living in the grouped mFUA, then it is given a score of 74. This corresponds to the blue parts in Figure 4.
4. Special Case - Bridges: if one mFUA is matched to two or more non-adjacent clusters of administrative areas (because the administrative units between the two clusters did not reach the matching threshold), then a bridge needs to be determined to connect two parts of the mFUA. This is done by identifying administrative areas that connect the two or more clusters with a minimal additional administrative area. If the area of the bridge is less than 50% of the original mFUA's area, then the bridge is integrated into the mFUA and given a score of 59.
5. Special Case - Holes: for each mFUA, if there are administrative areas that lie entirely within it and are not identified as part of that mFUA, often because of irregular shapes of oFUA or the mFUA has an enclave that does not touch the oFUA, then it is identified as holes and integrated into the mFUA with a score of 68. If the administrative area is already assigned to a different mFUA and if it is the only administrative area of that mFUA, then this mFUA is merged into the mFUA mentioned before and given a score of 58.

To allow differentiation of space in terms of urban density within medium to large-sized FUAs, we determine the DoU of each administrative area in the mFUA, by calculating the number of people living in each DoU cluster level for each geographical unit and find the DoU level within which more than 50% of the population live in. In the case that none of the DoU levels has more than 50% of the population living in it, the level is determined by accumulating the population from the lowest level upward, until reaching a population that is or greater than half of the population of the geographical unit, and this level will be given to that geographical unit. This will facilitate both national and international studies on urbanisation, and related studies that use urbanisation as a factor for consideration.

5 Results

After applying the aforementioned algorithm to the fifth dataset, the results indicate that out of 3328 oFUAs in these 40 countries, 1177 of them are successfully identified as mFUAs, meaning that 35% of all oFUAs are matched. When looking at country-specific matching levels, the results vary considerably, from more than 90%

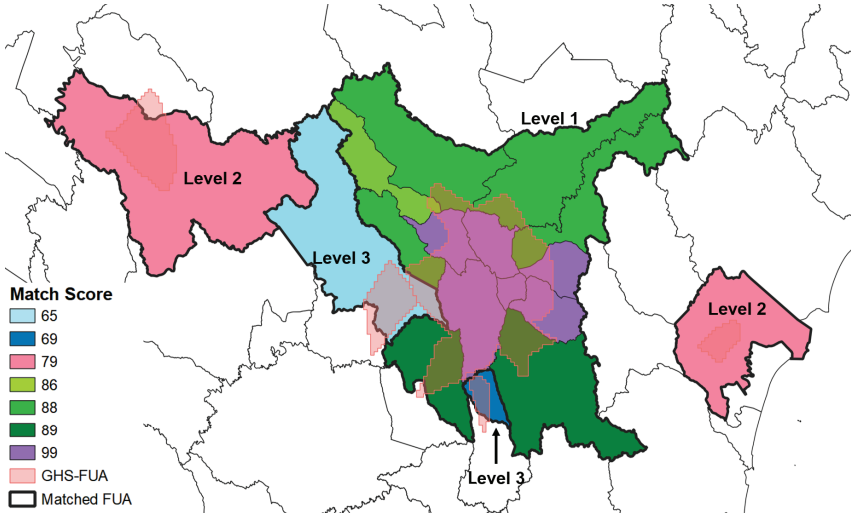

Figure 4: Matching Results Illustration

of oFUAs in Cuba, Panama, Uruguay, Chile, Brazil, Mexico and Bolivia being matched to mFUAs, to less than 10% in Uganda, China, Kenya, Burkina Faso and Mali. This variation is mainly due to the average area of each country's geographical units. Figure 5 shows exactly the correlation between the average area of geographical units on the x-axis and the percentage of matched oFUAs (number of mFUA / number of oFUA) for each country. The countries with only one mFUA are removed outliers. Figure 5 indicates that the smaller the average area of geographical units, the higher the percentage of matches. This is because smaller geographical units can better take into account the irregularity of oFUA boundaries, whereas many larger units are often much bigger than the intersecting oFUAs, making it impossible to classify those oFUAs using geographical units.

Of the 1177 matched FUAs, 284 are level 1 matches, 678 are level 2 matches and 215 are level 3 matches. Figure 6 shows an overview of all 1177 mFUAs. Visual inspection indicates that mFUAs under "bad" matches and "intermediate" matches are much more frequent, as well as occupying much bigger areas compared to "good" matches. When performing the analysis, it is recommended to focus mainly on "good" and "intermediate" matches, while treating "bad" matches with caution. This is because "good" and "intermediate" matches tend to represent the oFUAs reasonably well, whereas "bad" matches tend to overestimate the population of FUAs. Figure 7 has been generated from an earlier version of the results to demonstrate this statement. The x-axis shows the population of mFUAs, and the y-axis shows the population of oFUAs. The FUAs with labels "good" and "intermediate" matches are close to the reference line, whereas those of "bad" matches are below the reference, indicating that mFUA population is often much greater than the corresponding oFUAs.

The use of the mFUA is preferable to the use of the city proper, as the city proper is often not an accurate representation of cities and often leads to an underestimation of the population, leading to significant misunderstandings when studying the urban population. A visual example of this phenomenon is shown in Figure **??**, the oFUA of Rio de Janeiro. The city proper of Rio de Janeiro consists only of the southernmost polygon with the same name, whereas the oFUA is much larger than the city proper. Similarly, Figure 8 shows the difference between the population of the city proper and mFUAs for all countries using the previous version of the results. The x-axis corresponds to the mFUA population, and the y-axis shows the population of the corresponding city proper, for all cities where the city proper could be found. The graph shows that an overwhelming number of mFUAs have much larger populations than the corresponding city proper, especially for Asian and European cities. At the same time, larger mFUAs appear to have higher population differences than smaller mFUAs. This further confirms the need to use the FUA definition to delineate city boundaries, rather than government definitions. Note that Figure **??** and Figure **??** will be updated and studied further using the GADM dataset covering all countries in the following studies.

Figure 9 gives an overview of the DoU classification of all geographical units in the fifth dataset. Out of 10064 geographical units, 18% are classified as urban centres, 17% as dense urban areas, 6% as semi-dense urban areas, 20% as suburban areas, 23% as rural areas, 15% as low-density rural areas and 3% as very low-density rural areas. This set of results can be generated every five years as GHSL publishes data every five years since 1975

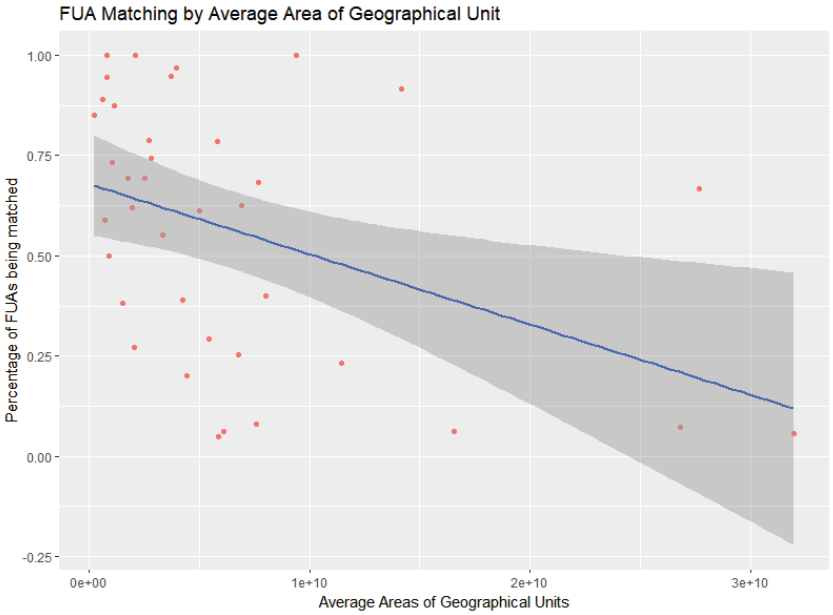

Figure 5: FUA Matching by Average Area of Geographical Unit

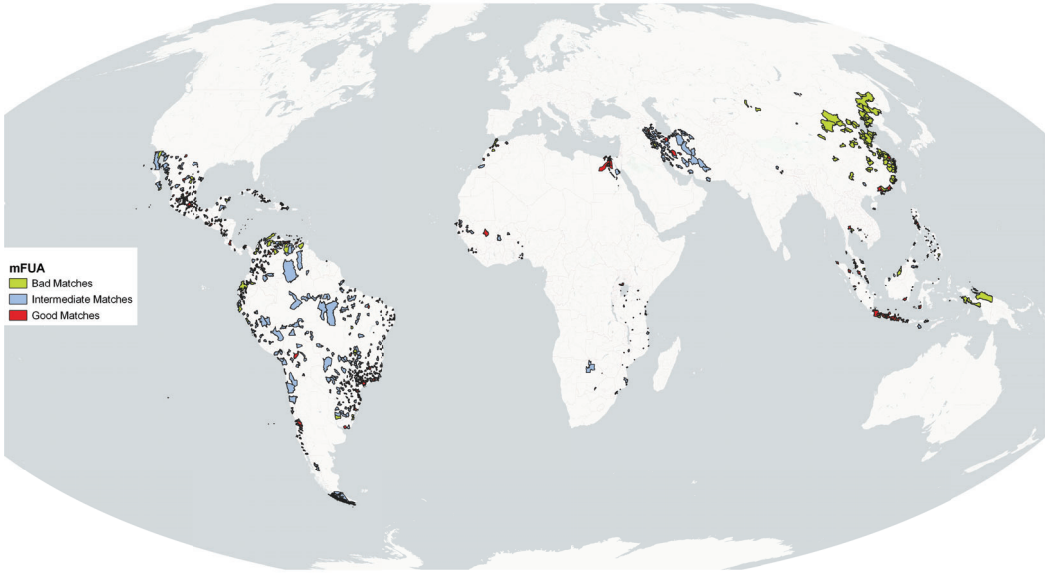

Figure 6: mFUAs in the Fifth Dataset, as of 2015

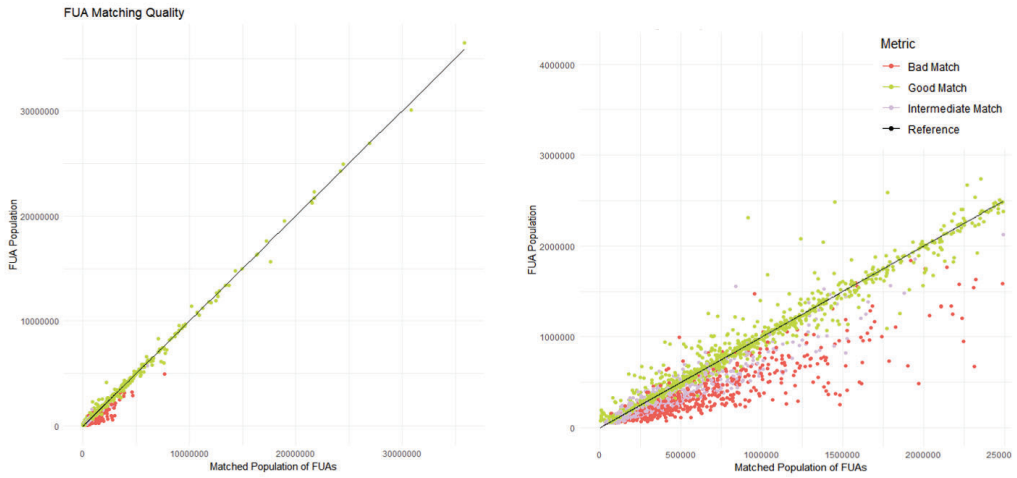

Figure 7: FUA Matching Quality

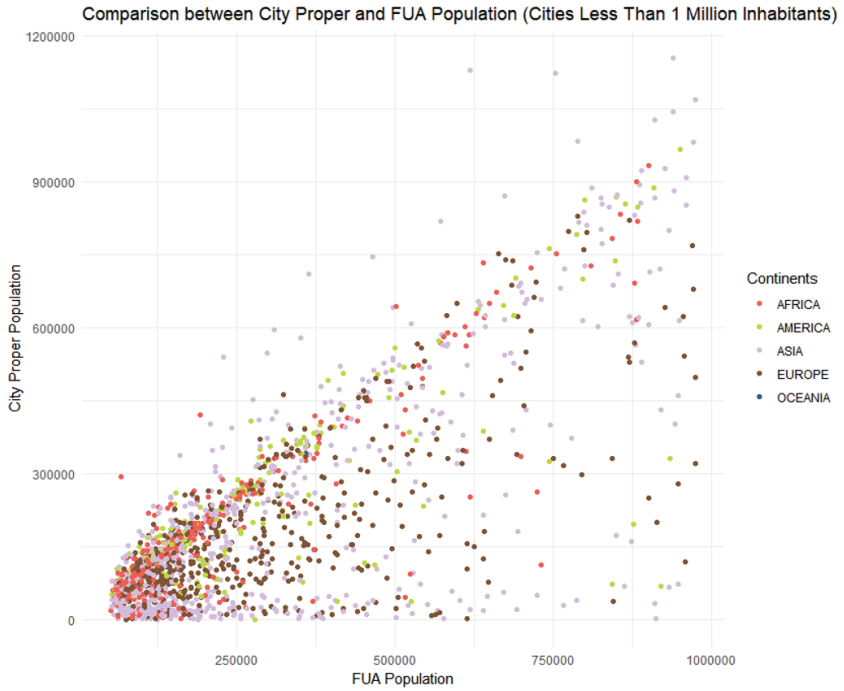

Figure 8: City Proper versus mFUA - a Comparison

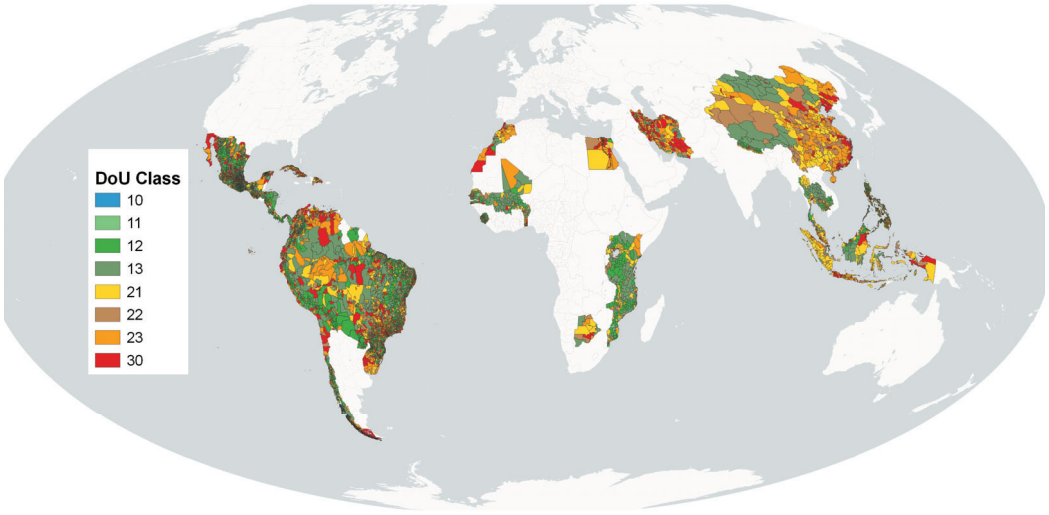

Figure 9: DoU Classifications for All Geographic Units in the Fifth Dataset, as of 2015

(European Commission, 2023). This is useful for monitoring the urbanisation process of specific regions, for comparing urbanisation across regions, and for conducting urban studies within cities. This further illustrates the richness of the GHSL data and its potential application to the research community.

For the next communication, the data will be generated for all countries around the world based on GADM shapefiles. The corresponding analysis will also follow up.

6 Conclusion

The GHSL FUA dataset is an advantageous definition for delineating city boundaries because it is a more realistic representation of city boundaries than the traditional definition of a city. In addition, it is defined using a globally harmonised dataset, making it suitable for internationally comparable studies as well as for national studies where local governments do not define the actual city boundary. The algorithm presented in this article further extends the use of the FUA dataset to include datasets that are aggregated by the geographical units of each country, such as the microcensus data from IPUMS, where individual-level data have been grouped into national geographical units to protect the privacy of individuals. This algorithm is also useful for those who wish to study cities and urbanisation on a large scale, where the manual definition of FUA boundaries is too costly and complicated. This methodology could be applied to a wide range of studies, from urban demography to urban planning and economics.

References

European Commission. (2023). *GHSL data package 2023* (tech. rep.). Publications Office. Luxembourg. Retrieved May 23, 2023, from https://data.europa.eu/doi/10.2760/098587

Eurostat. (2021). *Applying the degree of urbanisation: A methodological manual to define cities, towns and rural areas for international comparisons : 2021 edition.* Publications Office. Retrieved September 12, 2023, from https://data.europa.eu/doi/10.2785/706535

GADM. (2023). Database of Global Administrative Areas. Retrieved October 13, 2023, from https://gadm.org/index.html

Loibl, W., Etminan, G., Gebetsroither-Geringer, E., Neumann, H.-M., Sanchez-Guzman, S., Loibl, W., Etminan, G., Gebetsroither-Geringer, E., Neumann, H.-M., & Sanchez-Guzman, S. (2018). Characteristics of Urban Agglomerations in Different Continents: History, Patterns, Dynamics, Drivers and Trends. In *Urban Agglomeration*. IntechOpen. https://doi.org/10.5772/intechopen.73524

Minnesota Population Center. (2019). Integrated Public Use Microdata Series, International: Version 7.2. https://doi.org/10.18128/D020.V7.2

Moreno-Monroy, A. I., Schiavina, M., & Veneri, P. (2021). Metropolitan areas in the world. Delineation and population trends. *Journal of Urban Economics*, *125*, 103242. https://doi.org/10.1016/j.jue.2020.103242

OECD. (2012). *Redefining "Urban": A New Way to Measure Metropolitan Areas*. Organisation for Economic Co-operation; Development. Retrieved September 12, 2023, from https://www.oecd-ilibrary.org/urban-rural-and-regional-development/redefining-urban_9789264174108-en

United Nations Department of Economic and Social Affairs. (2006). *United Nations Demographic Yearbook 2000*. United Nations. https://doi.org/10.18356/875dad7e-en-fr