Journal of Urban Economics xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Urban Economics

journal homepage: www.elsevier.com/locate/jue

Metropolitan areas in the world. Delineation and population trends

Ana I. Moreno-Monroy [a], Marcello Schiavina [b], Paolo Veneri [c,*]

[a] OECD Centre for Entrepreneurship, SMEs, Regions and Cities, Honorary Associate, Geography and Planning Department, University of Liverpool, 2 rue André-Pascal, 75016 Paris, France
[b] European Commission, Joint Research Centre, Via E. Fermi, 2749, 21027 Ispra VA, Italy
[c] OECD Centre for Entrepreneurship, SMEs, Regions and Cities, 2 rue André-Pascal, 75016 Paris, France

ARTICLE INFO

JEL classification:
R11
R1
R14

Keywords:
Cities
Metropolitan areas
Functional urban areas
Suburbanisation

ABSTRACT

This paper presents a novel method to delineate metropolitan areas – or functional urban areas (FUAs) – in the entire world and assesses their population trends. According to the definition developed by the OECD and the European Union, FUAs are composed of high-density urban centres with at least 50 thousand people plus their surrounding commuting zones. The latter represent the urban centres' areas of influence in terms of labour market flows. The proposed method combines a functional and a morphological approach to overcome the dependency on travel-to-work data to define commuting zones and allow a global delineation. It relies on a probabilistic approach and the use of population and travel impedance gridded data across the globe. Results show that around 3.9 billion people, making up 53% of the world population, live in 8,790 FUAs, out of which 17% live in their commuting zones. Between 2000 and 2015, population growth was higher in larger FUAs, highlighting a general trend toward higher concentration of the metropolitan population. Commuting zones grew faster than urban centres, though with heterogeneous patterns across world regions, income levels and metropolitan size.

1. Introduction

Understanding socio-economic and demographic dynamics of metropolitan areas requires a careful delineation of metropolitan boundaries. Metropolitan areas are composed of densely inhabited urban centres plus their surrounding and interconnected lower-density areas. A major reason behind the need to delineate metropolitan areas is that official data and information at that scale generally refer to administrative or legally-defined regions. The latter tend to adapt slowly to rapid changes in population and economic activities in space, yielding a persistent or even increasing misalignment between 'legal' and 'economic' boundaries. Moreover, large differences across countries in the size and structure of local administrative units seriously affect cross-country comparisons and represent an obstacle to robust worldwide evidence on the features of urbanisation and its consequences.

Metropolitan areas' delineation generally adopts functional approaches, relying on commuting ties between local units (Duranton, 2015; Bosker et al., 2019). Such methods are likely to be the most accurate to delineate metropolitan areas, but the lack of commuting data in many countries limit a global and consistent delineation. Many countries have also delineated their own metropolitan areas for planning or statistical purposes (e.g. the Metropolitan Statistical Areas in the United States or the Census Metropolitan Areas in Canada). In the same vein, the Functional Urban Areas (FUAs) developed by the OECD and the European Union provide a consistent definition of metropolitan areas already applied to OECD and European countries (OECD, 2012; Dijkstra et al., 2019) based on an aggregation of local units (i.e. municipalities, small local jurisdictions, etc.). According to the EU-OECD definition, FUAs consist of high-density urban centres ('cities') and their surrounding areas of influence in terms of travel-to-work flows ('commuting zones'). In the remainder of the paper, the terms FUAs and metropolitan areas are used as synonyms.

While the scarcity of comparable data sources can limit global delineations of urban centres and metropolitan areas, recent studies rely on geo-spatial data generated from satellite imagery. This is the case for studies based on nighttime lights (Zhang and Seto, 2011; Ch et al., 2019; Dingel et al., 2019; Harari, 2019; Ellis and Roberts, 2016), built-up land cover classifications (Baragwanath et al., 2019), and a combination of built-up area and population estimates (Freire et al., 2018).

Our work contributes to this literature by proposing a method to uniquely identify commuting zones around urban centres. We start from considering urban centres as clusters of contiguous grid cells of high-density population, based on the definition developed by Dijkstra and Poelman (2014). We approximate commuting zones around each urban

* Corresponding author.
E-mail addresses: ana.morenomonroy@oecd.org (A.I. Moreno-Monroy), marcello.schiavina@ec.europa.eu (M. Schiavina), paolo.veneri@oecd.org (P. Veneri).

https://doi.org/10.1016/j.jue.2020.103242
Received 6 September 2018; Received in revised form 10 February 2020
Available online xxx

centre by relying on the estimated probability that one-km^2 grid-cells outside urban centres belongs to a metropolitan area.

The estimation is performed through a logistic regression model, which is trained using information on actual EU-OECD FUA (baseline) boundaries in 31 countries with available data. The estimation uses about 0.5 million one-km^2 cells with at least 300 inhabitants in baseline countries. The dependant variable takes the value of one if the cell falls within a baseline boundary and zero otherwise. Following previous work on the area of influence of cities (Uchida and Nelson, 2011), the predictors include the distance of the cell to the closest urban centre, the size of the urban centre, the cell population and country-level characteristics. We then use the model parameters to obtain estimated probabilities for around 2.5 million cells in and outside baseline countries. In order to define which cells belong to FUAs and which do not, we compare the estimated probabilities to optimal thresholds calculated by world region. In the same way as other satellite data-base definitions (Ch et al., 2019; Baragwanath et al., 2019), the resulting estimated FUAs are fully independent from country-defined local jurisdictions boundaries, thus maximising cross-country comparability.

Conceptually, our method comes close to studies estimating market potential areas around urban centres, including the agglomeration index method proposed by Uchida and Nelson (2011). However, our method differs in two important aspects. First, instead of imposing a predetermined distance threshold (i.e. 60-min travel time by car), we use a data-driven approach to determine the appropriate distance threshold for each urban centre in every direction by relying on estimated probabilities at the cell level. Second, in Uchida and Nelson (2011), market potential areas can overlap with other urban centres or the market potential areas of other urban centres. In contrast, our approach defines commuting zones that are unique to each urban centre, mainly because we want to obtain an estimate of population in commuting zones.

One important feature of our approach is that it delineates agglomerations of people rather than agglomerations of human activities, such as built-up areas or nightlights. Daytime satellite data captures a larger array and variety of human settlements than nighttime lights-based methods (Baragwanath et al., 2019), and may be more suited to capture differences in development levels and physical structures because built-up areas per capita likely increase with the level of economic development. Besides, nighttime lights-based methods involve a trade-off between under-identification of smaller urban settlements at the cost of "exploding" large urban areas to unrealistically large sizes because they rely on exogenous radiance thresholds (Baragwanath et al., 2019).

Our method relies instead on data from existing metropolitan areas across world regions, which allows us to conduct extensive sensitivity analyses to understand the consequences of necessary assumptions when attempting a global definition. While we train our model using information on mostly developed countries to make global estimations, our results come close to those of recent studies applied to both developed and developing countries (e.g. Brazil, China, India and the United States) relying on nighttime light and built-up land (Ch et al., 2019; Baragwanath et al., 2019). Moreover, our metropolitan area boundaries for large cities in Indonesia come close to boundaries obtained by Bosker et al. (2019) using the commuting flows method proposed by Duranton (2015).

By distinguishing urban centres from their respective commuting zones, our method allows us assessing world patterns of metropolitan population growth as well as of intra-metropolitan dynamics, including suburbanisation. We define suburban population as the population in commuting zones, and suburbanisation as the growth of population in those areas.

The method allows the delineation of 8,790 metropolitan areas (FUAs)[1] based on 10,082 urban centres in 168 UN-recognised countries. According to such results, 53% of the world population – about 3.69 billion people – lived in FUAs in 2015, out of which 17% lived in commuting zones. North America hosts the largest percentage (72%) of people in FUAs, followed by Latin America (63%). The concentration of metropolitan population in a few large FUAs is highest at intermediate levels of development and show an inverse U-shaped relationship with income per capita. On the other hand, the proportion of FUA population in commuting zones increases with income levels and is largest in high-income countries (31%).

Between 2000 and 2015, population growth was relatively faster in FUAs than in other areas, with relatively faster growth in larger metropolitan areas. Yet, one fifth of FUAs worldwide declined or stagnated. In terms of intra-metropolitan dynamics, the proportion of suburban population increased practically everywhere, although in absolute terms population in commuting zones remains significantly lower than in urban centres. Growth in the commuting zones coexisted with urban centres declines in 17% of FUAs, although in most cases urban centres and commuting zones showed a consistent change in population. Overall, 70% of FUA population growth occurred through densification of existing urban centres, although that proportion drops to only 56% in high-income countries, where suburbanisation and urban centres' expansion characterised more prominently metropolitan growth patterns.

The remainder of the paper is organized as follows. Section 2 outlines the delineation question and describes data sources and pre-processing. Section 3 outlines the empirical method to select, estimate and validate the econometric model, the procedure to delineate boundaries, and external validity checks using the cases of Brazil and Indonesia. Section 4 presents the results of the metropolitan area delineation by documenting and discussing key facts on population dynamics across and within FUAs. The last section concludes.

2. Problem definition and data preparation

2.1. Data sources

Our aim is to classify the earth's surface area into areas that belong to a FUA and those that do not. For this, we rely on three main sources of data: 1) global measures of urban centres and density; 2) global measures of travel time; and 3) existing FUA boundaries (see Annex A for a detailed description of the data).

The satellite-derived Global Human Settlements Layer (GHSL, Pesaresi et al., 2016) provides built-up and population distribution grids, and a settlements classification according to their density. This population layer is the basis for the settlement layer (GHS-SMOD, Florczyk et al., 2019; Pesaresi et al., 2019) that implements the global definition of cities proposed by Dijkstra and Poelman (2014). This harmonized definition allows the comparison of different cities extents or populations across the globe.

To obtain comparable measures of travel time between any two grid cells, we rely on the global travel impedance grid (Weiss et al., 2018). This grid represents time associated with moving through grid cells, quantified as a movement speed within a friction grid. Information on roads (fastest type takes precedence, with speeds obtained from Open Street Maps tables), railroads, water bodies and movement over land is used to characterize each grid cell. The unit of measurement is minutes required to travel one kilometre. See Weiss et al. (2018) for more details.

Finally, baseline metropolitan area boundaries – the EU-OECD FUAs boundaries available in 31 countries – define commuting zones based on the intensity of commuting to urban centres from areas outside them, following the method outlined in OECD (2012) and Dijkstra et al. (2019). The boundaries consolidate administrative borders of local units in each country.

[1] A geopackage with the full set of metropolitan area boundaries and their corresponding area and population in urban centres and commuting zones (2015) is freely available for download at: https://ghsl.jrc.ec.europa.eu/stg19/a2019/datasets.php. See Schiavina et al. (2019).

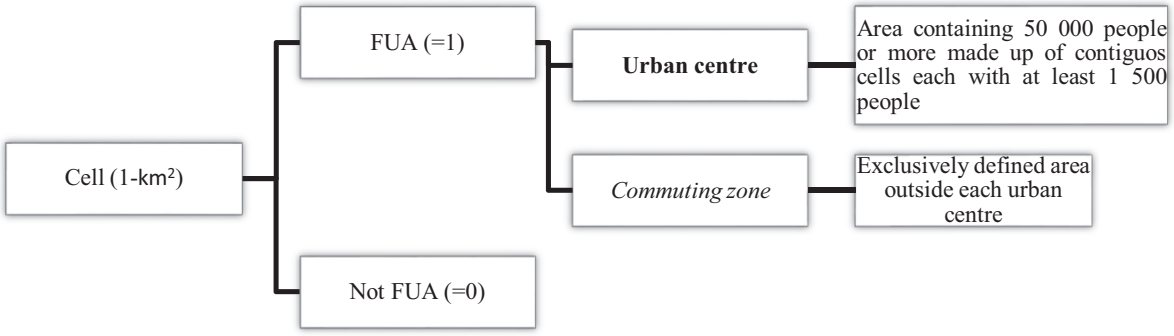

Fig. 1. Classification of cells and hierarchy of spatial aggregations.
Source: Elaboration based on Dijkstra and Poelman (2014).

2.2. Problem definition

Based on the GHS-SMOD layer and consistently with the settlement layer definition (Dijkstra and Poelman, 2014; Florczyk et al., 2019), the world land surface can be divided into two types of cells. The first type are high-density cells that make up urban centres. Urban centres are clusters of contiguous high-density cells (with at least 1,500 people per km^2 or at least 0.5 km^2 of built-up area) which altogether contain at least 50,000 inhabitants (Fig. 1).[2] Cells within urban centres are by construction part of a FUA and therefore do not need to be classified. The second type are inhabited cells with moderate (at least 300 people per km^2) or low density (less than 300 people per km^2).[3] Any cell outside urban centres may or may not be part of a commuting zone.

With a comprehensive global definition of urban centres, what is missing to obtain FUAs at a global level is defining the boundaries of *commuting zones* for each urban centre. It is not possible to extend the baseline FUAs delimitation method globally because we lack the necessary commuting flows data. However, we can use baseline FUAs to estimate the probability that each cell within each country is part of a commuting zone of an existing urban centre.

Our empirical strategy is to train a classification model using baseline FUAs boundaries to decide which cells are part of commuting zones and which ones are not. We can then evaluate the cell-level probabilities against an optimal probability threshold to determine which cells belong to FUAs (i.e. those = 1) or not (those = 0) (Fig. 1).

2.3. Data preparation

Before proceeding with the model estimation, we need to undertake two data preparation steps: 1) subset the group of cells that need to be classified; and 2) assign them to urban centres.

Regarding the first step, we use a subset of cells with population above a single predetermined threshold to determine the boundaries of commuting zones, instead of using all populated cells. This strategy greatly reduces the number of cells that need to be classified globally. While the choice of threshold could be made based on model performance metrics for different options, this strategy is computationally intensive[4] since the number of cells increases greatly at low-density values (and will ultimately approximate the earth's area at the extreme case when all cells are included). As an alternative, we use a threshold of 300 people which corresponds to the threshold officially applied by Eurostat to define medium-density cells in the Degree of Urbanisation definition (Dijkstra and Poelman, 2014). From here onwards, when the term "cell" is used, it will refer to a cell with a population of at least 300 people, unless otherwise stated.

Regarding the second step, assigning every cell to a unique urban centre allows us to get non-overlapping commuting zones. We assign unequivocally each cell to an urban centre following a simple rule. In baseline FUAs, local areas are assigned as part of the commuting zone of the urban centre with the largest commuting intensity amongst all urban centres in the country, provided the percentage of people commuting from the local area is above a certain threshold. Without commuting flows data, a reasonable assumption is to assign each cell to the closest urban centre (i.e., assume that the closest centre has the highest probability amongst all other urban centres). In baseline FUAs, cells inside FUAs (=1) are already assigned unequivocally to an urban centre and a country, while all other cells outside FUAs (=0) are only assigned to a country. This means that for the baseline case, we apply the procedure of assigning cells to the closest urban centre only to cells outside baseline FUAs. In all other cases, we assign all cells outside urban centres to their closest urban centre (i.e., the urban centre with the smallest travel time).

3. Empirical strategy, implementation and validation

After assigning and selecting the subset of cells that will be the basis to construct commuting zone boundaries for all urban centres, there is at least a couple of major choices to make regarding which cells will be ultimately categorised as part of a FUA. The first concerns the specification of the classification model trained for 31 countries and whose parameters are used to estimate FUAs' boundaries globally (Fig. 2, step 1). The second concerns the probability threshold that defines which cells are considered part of a FUA or not (Fig. 2, steps 2 and 3).

3.1. Model specification and selection

Our model selection problem consists of verifying the relevance of the predictors suggested by theory and previous studies, and finding a probability distribution that balances realism with parsimony. In other words, we want to come as close as possible to the true data generating process without overfitting the data.

Previous literature on the extent of cities emphasises the role of agglomeration economies generated by thick local labour market and enhanced learning and matching mechanisms (Duranton and Puga, 2004). Building on Uchida and Nelson (2011) and their agglomeration index

[2] Urban centres that cross country borders are split by country, and only those with at least 50 thousand people are kept. This is done because the estimation uses country-level data and relies on baseline FUAs defined within country boundaries.

[3] Including uninhabited cells (e.g. water bodies, parks, etc.) which are not considered for the classification problem but may be part of FUAs (e.g. a park within an FUA).

[4] The implemetation for a given specification using cells with more than 300 people takes several days running in high-capacity servers, making unfeasible the task of running global versions of the model for multiple specifications for all cells.

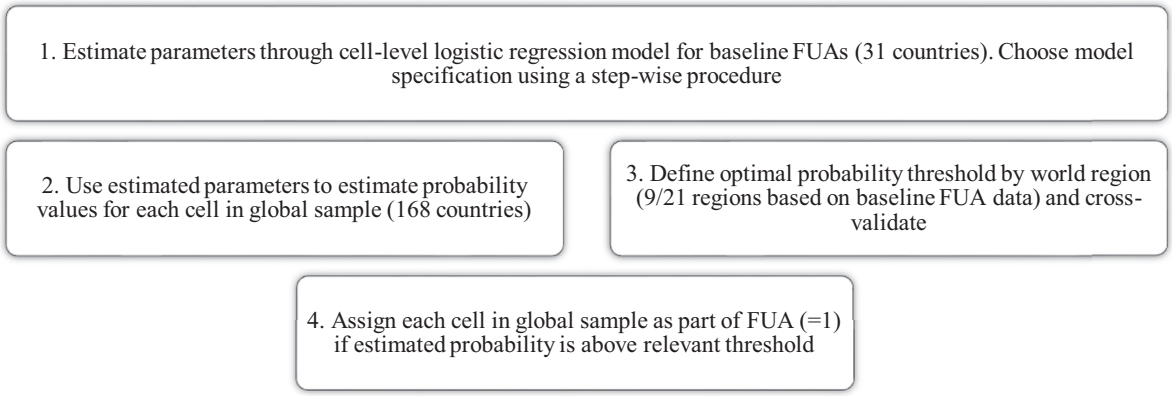

Fig. 2. Classification steps.

identification, three key indicators can capture the sources of agglomeration economies and can be framed under a gravitational approach: the population size and the density of the urban centre on the one hand, and the travel time to the urban centre on the other. While size and density are linked with agglomeration benefits in urban centres, low distances ensure surrounding lower density areas can access these benefits.

Translating these arguments into an econometric specification leads to the following general model to delineate FUA boundaries at the cell level based on cell, urban centre and country characteristics:

$$FUA_{ijC} = \beta_0 + \sum_{k=1}^{p} \beta_k dist_{ijC}^k + \sum_{k=1}^{p} \gamma_k size_{ijC}^k + \sum_{k=1}^{p} \varphi_k size_{jC}^k + \sum_{k=1}^{p} \omega_k GDP_C^k + \sum_{k=1}^{p} \pi_k cars_C^k + \in_{ijC} \quad (1)$$

where FUA_{ijC} is a dummy variable that takes the value of one if cell i located in country C and linked to urban centre j falls within FUA boundaries and zero otherwise; $dist_{ijC}$ is the travel time between cell i and urban centre j; $size_{ijC}$ is the size of the cell (measured by population); $size_{jC}$ is the size of urban centre j (measured by population, area or nightlight); GDP_C is GDP per capita; $cars_C$ is the number of vehicles per 1,000 inhabitants; and $\in_{ijC}$ is an error term; k is the degree of each predictor in the summation (from 1 to p) with its own coefficient. Annex A describes the data and its sources, and Table B.1 in Annex B shows summary statistics for the variables considered. To capture meaningful differences in car usage across countries, we would need to consider road capacities and/or a measure of transit (congestion) in addition to road provision already captured in the global impedance matrix. Unfortunately, this information is not currently available at a global scale in a comparable form.

The estimation method is a logistic regression with binomial distribution via nonparametric bootstrap with 100 repetitions and clustered standard errors at the FUA level.[5] The estimation relies on 466,361 observations across 1,287 urban centres in 31 countries: Colombia plus OECD countries, except those with no available FUAs (New Zealand, Israel, Turkey and Lithuania) and those with only one FUA (Luxemburg and Iceland). The proportion of cells within FUAs is 46.8%, while the reminding 53.2% are outside FUAs. All variables on the right-hand side are log-transformed as in other applications based on the gravity model (e.g. Ahlfeldt and Wendland, 2016; Goh et al., 2012).

The model selection relies on a stepwise model formulation that adds one predictor at the time. The procedure starts with the model with the smallest Bayesian Information Criteria (BIC) among those with a single predictor out of $dist_{ijC}$, $size_{jC}$, $size_{ijC}$, GDP_C and $cars_C$. At this step, we compare three different proxies for urban centre size: area, population and nightlight. The next step involves testing two predictors at the time, adding the square term of the already selected variable, and choosing the model with the lowest BIC value. Each time a specification is kept if all parameter estimates (including the intercept) have a significant p-value (<0.05).[6] Once this model is obtained, all possible interactions among the remaining predictors are tested and kept if they are statistically significant and improve the model performance. Annex B illustrates the procedure at step 1 and from step 5 onwards (results for other steps are available upon request). Tables B.2 and B.3 in Annex B shows the regression results.

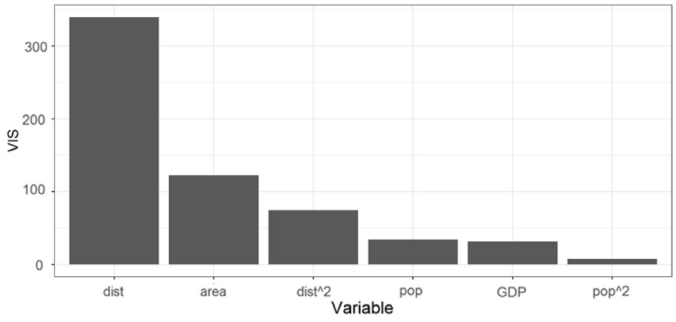

Fig. 3. Variable importance score.
Note: VIS = Variable Importance Score, dist = travel time, area = urban centre area, pop = cell population.

The stepwise procedure supports the following specification:

$$FUA_{ijC} = \beta_0 + \sum_{k=1}^{2} \beta_k dist_{ijC}^k + \gamma_1 size_{ijC} + \sum_{k=1}^{2} \varphi_k pop_{jC}^k + \omega GDP_C + \in_{ijC} \quad (2)$$

using population as proxy for urban centre size ($size_{jC}$). Extensive sensitivity analysis shows that the final delineation of FUAs' boundaries is robust to changes in the model specification (results available upon request). Annex C shows a comparison with a specification including random intercepts at the level of urban centres and countries. Fig. 3 summarizes the variable importance score (absolute value of the z-stats) of the estimated model. Travel time from each cell to the urban centre is the most powerful predictor of the probability of belonging to a FUA, followed by the size of the urban centre.

[5] Galdo et al. (2019) in their application for India also use bootstrapping across known urban areas.

[6] A caveat of stepwise procedure for model selection is that it is myopic, i.e., it only looks one step forward at any point, without any indication on the next variable to include.

3.2. Optimal threshold determination

After obtaining the model parameters, we can proceed to allocate cells with estimated probability higher than a given threshold to a FUA. However, we must first define a criteria to determine which threshold we will use. Because our aim is to maximize prediction performance, instead of establishing theoretically meaningful relationships between the explanatory variables and the outcome,[7] we determine an optimal probability threshold by maximising model accuracy (A) of predicted positives and negatives, as follows:

$$\max_t \; A\,(t) = (TP(t) + TN(t))/(TP(t) + TN(t) + FP(t) + FN(t)) \qquad (3)$$

where A is accuracy, t is the optimal threshold value; TP are the truly predicted positives; TN the truly predicted negatives; FP the falsely predicted positives; and FN the falsely predicted negatives. A larger share of falsely predicted positives leads to higher false positive error, or acting when action was not warranted (Type I error). Similarly, a larger share of falsely predicted negatives lead to larger false negative error, or to not acting when action was warranted (Type II error). The results of this maximization problem will be influenced by how much we weight false positive versus negative error. In our case, larger false positive error may lead to FUAs that are too large (cells which should not have been included are included), and larger negative error may lead to FUAs that are too small (cells which should have been included are not included). As neither of these options is clearly desirable *a priori*, we make our decision on the appropriate weight based on model performance metrics described below.

Besides the positive versus negative error weights, another choice we need to make is whether we use a single optimal threshold for the world or calibrate optimal thresholds for each world region. The latter strategy would allow for more prediction flexibility in regions with different urbanisation patterns, geographies and levels of development, but we still need to determine empirically if it performs at least as well as a single world threshold.

To determine thresholds for each world region, we use baseline information when available (i.e., in 9 out of 21 United Nations, UN, world regions) and full model information otherwise. For regions with no baseline country information in the training set, a conservative alternative is to use the highest estimated threshold (0.75).[8] For each UN region, the optimal threshold is determined by maximising the accuracy (A) as specified in Eq. (3) for each region, assuming that false positive error has an equal weight than false negative error.

We select the final optimal thresholds based on performance across five criteria: 1) sensitivity, or the percentage of cells inside FUAs rightly predicted as such; 2) Specificity, or the percentage of cells outside FUAs rightly predicted as such; 3) Balanced Accuracy, or the number of correctly predicted positives/negatives over the total number in each class; 4) False negative error; and 5) False positive error. These statistics are calculated over the actual number of cells inside and outside drawn FUA boundaries, which differ slightly from the corresponding estimated probability values because additional cells may enter the FUAs as part of the boundary drawing process (see the next subsection and Table O1 in Appendix G).

The best performing model is the one with optimal thresholds by world region and equal weights for positive versus negative error. Annex D describes the optimal threshold method in detail, including robustness checks in comparison with different false positive/false negative error weights. Table 1 shows the performance of the selected model. In absolute terms, the model performs above what generally accepted to be a good to excellent performance.

Table 1
Performance results.

Performance metric	Value
Sensitivity	0.8396
Specificity	0.8584
Balanced Accuracy	0.8490
False Positive Error	0.1416
False Negative Error	0.1604

3.3. Functional urban areas delineation

We delineate FUAs by combining cells selected through the steps described in Fig. 2 into polygons. This delineation procedure implies small changes in the cells that are finally included in the boundaries, as described in detail for each country in Table O1 in Appendix G. The next subsection shows robustness results for larger distance thresholds.

In order to avoid unrealistically large urban centres in terms of surface and population (an issue in some highly dense countries such as Bangladesh and Egypt), we impose two simple rules: first, a FUA can only have one urban centre of half a million inhabitants or more (see Annex E for robustness analysis for different options of these thresholds). Second, we split urban centres with more than 20 million inhabitants and more than 2,500 km^2 if they have at least two hypercores. Hypercores are areas within large urban centres containing 1 million people or more made up of contiguous cells, each with more than 3,000 people.[9] This splitting applies only to the six largest urban centres[10] and allows very high-density areas within their FUAs to be considered as independent commuting destinations.

In the baseline case, two or more FUAs with urban centres located at 5 km from each other are merged into a single polycentric FUA. This merging procedure is not applied if the population of either urban centre is at least 500 thousand people. The final delineation results in 8,790 FUAs, based on 10,082 urban centres in 168 countries.

3.4. Comparison of alternatives with different merging distance thresholds

Conceptually, increasing the merging distance threshold should lead to more fragmentation, i.e., a larger number of FUAs located at close proximity from each other and a more even size distribution. Alternatively, decreasing the merging distance threshold should lead to less fragmentation but also to a more uneven size distribution, as it can result in larger FUA sizes at the top of the size distribution.

In order to understand the effect of increasing or decreasing the merging distance threshold, the results for two alternative cases are considered: 1) not merging FUAs unless the urban centres touch (i.e. setting the merging distance to 0 km); and 2) increasing the merging distance threshold to 10 km. In this comparison, we keep the population threshold at which urban centres are not merged constant (and equal to 500 thousand people). Annex E shows the sensitivity results for this threshold.

The rank-size rule states that there is an ordering in city sizes such that the second largest city is half of the size of the first, the third half of the size of the second and so on. Empirically, we test the relationship by regressing the log of city rank against the log of the city population for the baseline case (5 km), and the two comparison cases (0 km and

[7] The balanced proportion of FUA versus non-FUA cells in the sample ensures better prediction performance is not mechanically related to a higher or lower proportion of cells in the sample. See Chapter 3 of Géron (2017) for more details on our approach on training a binary classifier, selecting performance measures and cross-validation using machine learning techniques.

[8] The value of 0.75 comes close to to the one obtained by weighting false positive error 3 times higher than false negative error (0.74).

[9] Urban centre area is divided according to the Euclidean distance from each hypercore.

[10] These are: Guanghzhou, Jakarta, Tokyo, Shanghai, Dhaka and Kolkata.

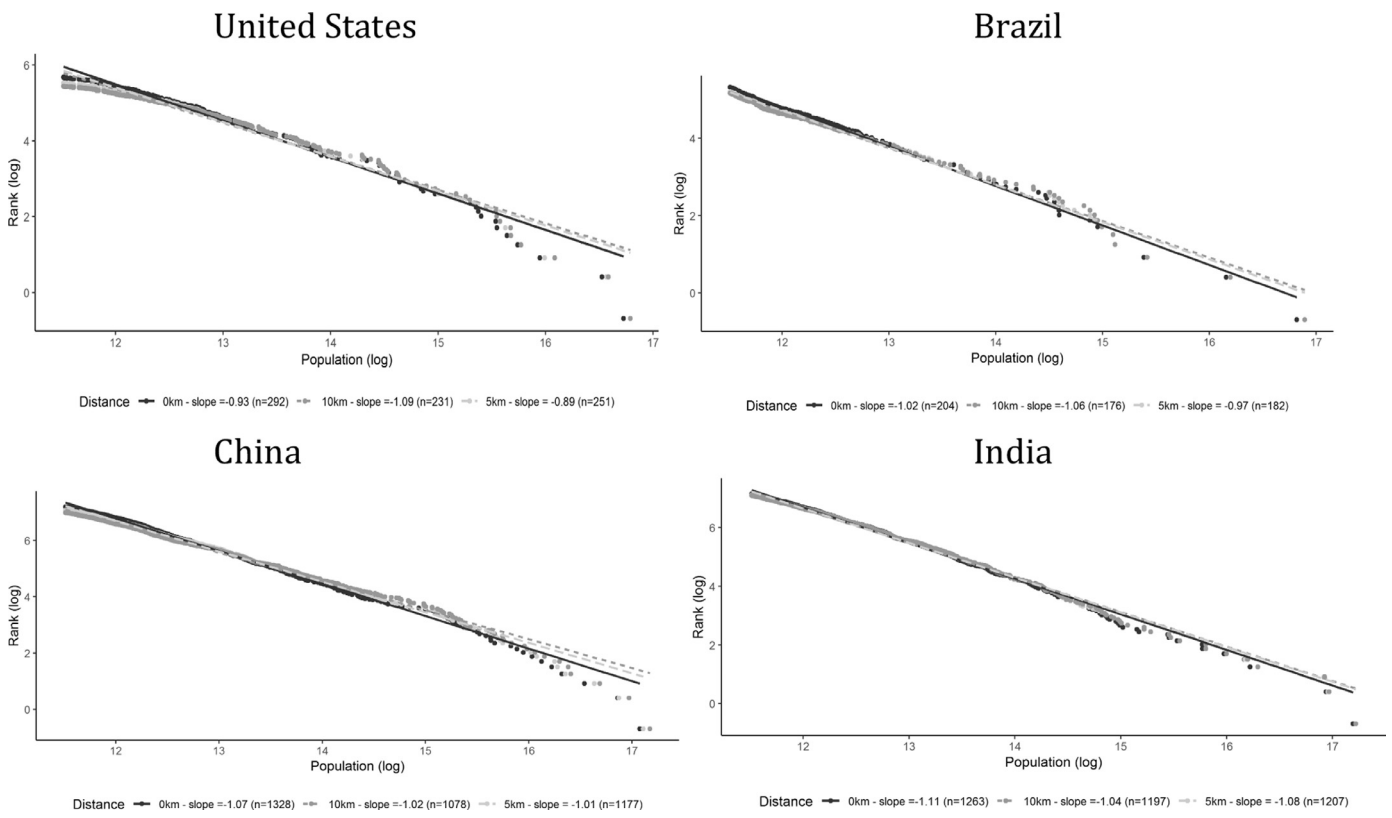

Fig. 4. Rank-size regression by selected countries, 2015.
Note: "Slope" corresponds to slope regression estimates of a regression on log(rank – 0.5) on log(FUA population) for 2015. All slope coefficients significant at the 1% level of confidence.

10 km).[11] The direction and magnitude of the shifts in the estimated rank-size relationship signals in which cases changes in the merging distance threshold significantly impact the urban-size distribution. For instance, more merging at the top of the city-size distribution leads to more primacy – or in terms of the rank-size rule estimation, to a smaller slope coefficient (Ch et al., 2019). As expected, a merging distance of 0 km results in smaller estimated slopes compared to the baseline case (5 km), indicating a faster decay at higher population values and a higher number of (smaller) FUAs. This is the direct result of less FUAs being merged. The differences in slope coefficients are nevertheless small.[12] On the other hand, the slope coefficients for the case of 10 km versus 5 km are larger, indicating a slower decay at higher population values and a smaller number of larger FUAs, because more FUAs have been merged into polycentric ones.

We can also use these rank-size results for United States, Brazil, China and India to compare our results with those of Chauvin et al. (2017), Dingel et al. (2019) and Ch et al. (2019).[13] Fig. 4 displays the rank-size regressions for United States, Brazil, China and India for the three merging distance options.

The shape of the rank-size fit and the regression coefficients we obtain are in line those reported in Ch et al. (2019) and Dingel et al. (2019). They confirm the relevance of the rank-size rule in describing the urban system of all four countries, as the estimated coefficients are all close to −1. Consistently with those studies, the smaller estimated coefficients (in absolute value) we obtain do not lend support to the hypothesis of non-linearities (possibly driven by primacy) in the rank-size fit for China and India initially proposed by Chauvin et al. (2017) using administrative areas.

Our slope coefficients for the rank-size regressions for the baseline case are within the range of results in the literature.[14] The shape of the rank-size plot and slope coefficients are remarkably similar in all cases except for India, especially when using the 5 km merging distance (i.e. the baseline case). This is the case even though we identify many more metropolitan areas compared to Ch et al. (2019), even when using a 10 km merging distance threshold.[15]

[11] We run rank-size regressions for countries with at least 20 FUAs (67 cases). All slope coefficients are within the expected range (∼ −1.4 to ∼−0.7), excect for Ethiopia (∼ −1.74). Empirical applications in the literature subtract 0.5 from the rank, following Gabaix and Ibragimov (2011)'s suggestion for improved estimates (Ch et al., 2019; Chauvin et al., 2017). For comparability this paper follows the same approach.

[12] In the comparison between 0 km versus 5 km the difference in slope coefficient is larger than 0.1 (in absolute value) for United Kingdom, South Africa, Canada and Egypt. In the comparison between 5 km and 10 km it is larger than 0.1 only for the United Kingdom.

[13] Chauvin et al. (2017) uses administrative areas to define metropolitan areas whereas Ch et al. (2019) and Dingel et al. (2019) use nigh-time lights. Their results are for 2010, whereas this paper's are for 2015. For comparability, FUAs with less than 100 thousand people are dropped in the four cases.

[14] Ch et al. (2019) report −0.81 (n = 201) for United States; (compare to −0.89 (n = 251) in baseline), −0.91 (n = 115) for Brazil (−0.97 (n = 182) in baseline results, and −0.96 reported by Soo (2014)), −0.96 (n = 534) in China; (−1.01 (n = 1,177) in baseline), and −0.92 (n = 344) in India; (−1.08 (n = 1207) in baseline). Dingel et al. report −1.178 (n = 1,264) in China (2010) and −1.044 (n = 465) in India (2011). Our results for the United States are also in line with Veneri (2016), who uses the EU-OECD FUA definition for OECD countries and finds a coefficient between −0.86 and −0.89 (n = 262).

[15] On the other hand, we identify a similar number of metropolitan areas compared to Dingel et al. for China (they obtain 1,264 cities in 2010 while we obtain 1,078 in 2015).

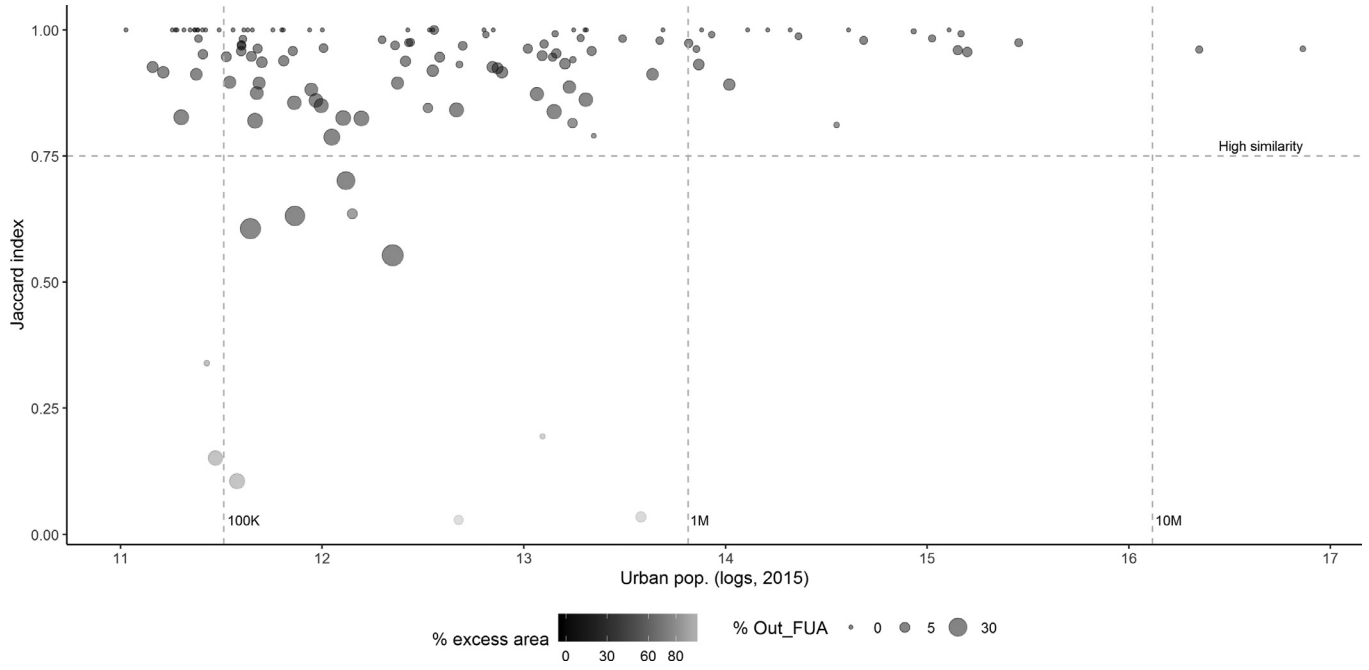

Fig. 5. Jaccard index results, Brazil.
Note: Out_FUA = Percentage of population in alternative boundaries not accounted for in administrative FUA boundaries; excess area = Percentage difference in area of administrative FUAs versus alternative boundaries. See Annex F for more details.

3.5. External validity

This section describes an out-of-sample comparison of estimated metropolitan boundaries for Brazil and Indonesia. The comparison relies on the Jaccard index (Bellefon et al., 2019; Bosker et al., 2019), which gives a measure of the overlap between two sets of boundaries by comparing the size of their intersection to the size of their union. The Jaccard index is complemented with other measures used to compare the likeness of the two sets of boundaries in each country. Annex F describes the Jaccard Index in more detail, as well as the procedure to translate FUA boundaries into administrative boundaries.

Brazil

The comparison boundaries for Brazil are drawn from the *arranjos populacionais* dataset. In a similar fashion as the EU-OECD definition of FUAs, this territorial definition relies on functional criteria based on commuting flows. Other metropolitan definitions for Brazil are limited to larger urban areas (e.g. 69 metropolitan areas), or are meant to be comprehensive of the whole territory without the adoption of any functional criteria (e.g. microregions, used in Chauvin et al., 2017). See Dingel et al. (2019) for a detailed discussion on the limitations of non-functional urban definitions for Brazil. See Annex F for details on the data preparation for the comparison.

Fig. 5 shows the comparison for the 128 arranjos against the 140 estimated FUAs subsequently adapted to local administrative boundaries. The median Jaccard index across 128 metropolitan areas is 0.96, indicating a very high level of concordance between the two sets of boundaries. The Jaccard index is equal to 1 (i.e. a perfect match) for 38 arranjos, and very high (>0.9) for 94 (in 128) cases, including the metropolitan areas of Rio de Janeiro (0.96), São Paulo (0.96) and Belo Horizonte (0.97). All metropolitan areas with more than 1 million people have Jaccard indices higher than 0.75. Mid-range values (0.5–0.7) occur in medium-sized arranjos, and low (<0.5) values occur in six cases. Administrative FUAs account for 6% less population compared to all 177 arranjos.

The lowest values of the Jaccard index occur in two cases: 1) when the comparison boundary is almost entirely covered by the large administrative FUA (e.g. Petropolis, Jaccard index = 0.02, covered by Rio de Janeiro); and 2) when FUA boundaries are intersected by a much larger arranjo (e.g. Jundai, Jaccard index = 0.03, intersected by São Paulo).

Indonesia

Unlike Brazil, there are no official metropolitan or functional urban boundaries with national coverage for Indonesia. A recent paper by Bosker et al. (2019) obtains metropolitan boundaries using different approaches, including the method based on commuting flows proposed by Duranton (2015). We compare boundaries obtained by Bosker et al. (2019) following this method using a 7% commuting population threshold. Such threshold is favoured by Bosker et al. (2019) because, among other reasons, it "generates a much larger number of separate areas, each consisting of only a few districts" while all the other "satellite data-based" approaches tend to agglomerate a larger number of districts in fewer of metro areas (Bosker et al., 2019).

Fig. 6 shows the results of the comparison. Out of 39 comparison areas, 2 are not covered by a FUA adapted to administrative boundaries. The median of the Jaccard index for 37 cases is 0.56, and the index takes the value of 1 for 10 cases. As Fig. 5 shows, for lower population values (below 1 million inhabitants), low Jaccard values arise because the comparison boundaries are larger than FUAs (thus FUAs do not account for a relatively large percentage of population in the comparison areas). In mid-range, low Jaccard values are due to smaller boundaries compared to FUAs (so population in the comparison boundaries is accounted for, but FUAs appear over-agglomerated). At high population ranges, both approaches give similar boundaries. The similarity is very high for the metropolitan areas of Jakarta (and they are also similar to official Jabodetabek boundary, population ~ 32 million), with a Jaccard value of 0.94, and for Bandung metro (second largest metro, 8.4 million), with a Jaccard value of 0.88 (see Fig. 7 for a visual of the boundaries in Java).

4. A global picture of metropolitan areas

By applying the method outlined in Section 3 we identify 8,790 FUAs across 168 countries, covering a surface of nearly 2.5 mil-

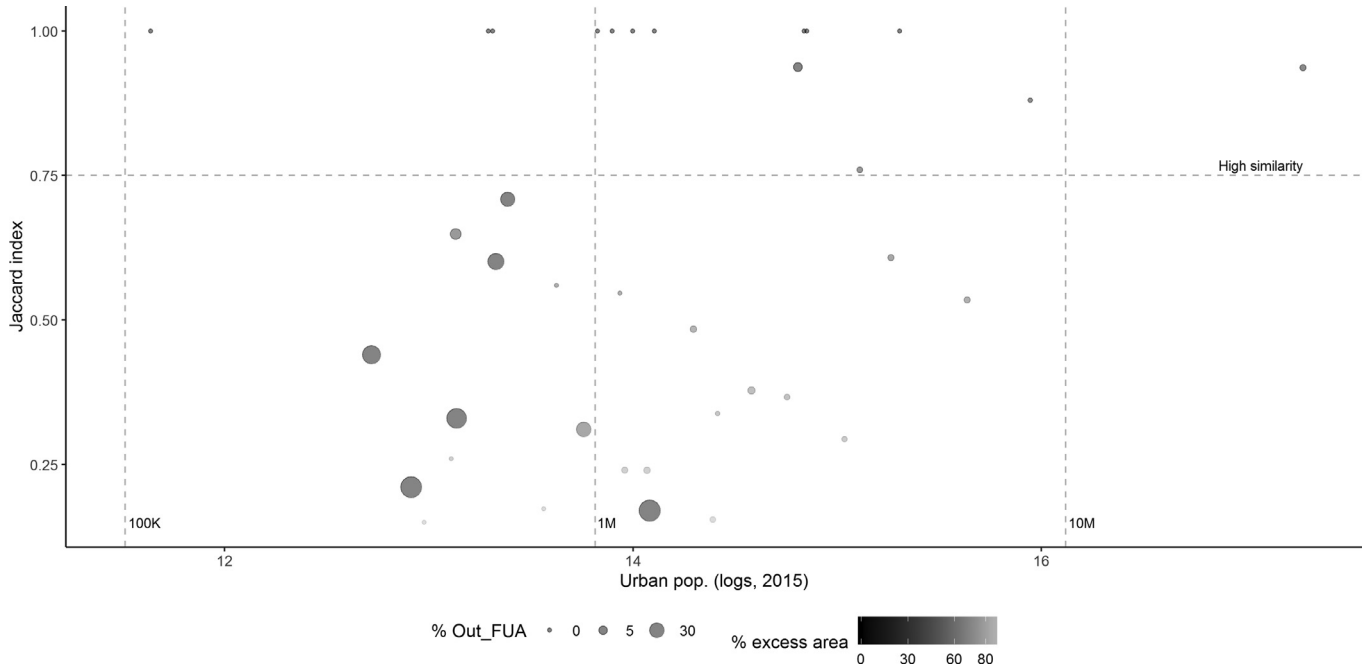

Fig. 6. Jaccard index results, Indonesia.
Note: Out_FUA = percentage of population in alternative boundaries not accounted for in administrative FUA boundaries; excess area = percentage difference in area of administrative FUAs versus alternative boundaries. See Annex F for more details.

Fig. 7. Java comparison of metro politan (metro) areas obtained by Bosker et al. (2019) using a 7% commuting threshold and FUAs adapted to administrative borders (admin. FUAs).
Source: Open Street Maps, Bosker et al. (2019) and own calculations.

Table 2
Functional urban areas by income level and world region, 2015.

	# FUAs	FUA population (thousands of persons)	FUA share (as% of total population)	Urban centres share (as% of total population)	Commuting share (as% of FUA population)
Income level					
Low income	842	205,113	33.2	32.0	3.6
Lower-middle income	3,696	1,455,400	49.6	44.4	10.4
Upper-middle	3,148	1401,400	55.2	45.5	17.7
High income	1,104	802,318	70.1	48.7	30.6
World region					
Africa	1,757	517,445	44.5	41.5	6.8
Asia	4,939	2,264,670	52.1	44.4	14.9
Europe	893	426,795	58.0	38.8	33.0
Latin America & Caribbean	868	375,176	62.5	54.7	12.5
North America	295	258,179	72.2	48.1	33.4
Oceania	38	22,054	58.7	42.5	27.5
World	8,790	3,864,319	53.4	44.4	16.9

lion km^2.[16] The United States has the largest FUA area coverage (570,477 km^2), followed by China (386,993 km^2) and India (141,244 km^2). The remainder of this section provides a set of key facts on the basic characteristics of metropolitan areas, as well as on their population dynamics across world regions and income groups.[17] We analyse population dynamics by keeping FUA boundaries constant at the time of delineation (2015) and use population levels in 2000 and in 2015, two available years in our global population data.

4.1. The importance and characteristics of metropolitan areas

Fact 1: More than half of the world population live in functional urban areas, with the largest share in North America (72%) and the lowest in Africa (44%).

With about 3.9 billion people in 2015, FUAs account for 53% of the world population. As expected, such a proportion changes with income levels, ranging from over 70% in high-income countries to 33% in low income countries (Table 2). The Americas have the largest share of people living in FUAs, followed by Europe and Asia, while Africa has the smallest share. With 36.4 million people, Tokyo, Japan is the world's largest metropolitan area, followed by Delhi, India.

Population in FUAs approximates the amount of dwellers whose connections, lifestyle and mobility is likely to be urban, including both people living in high-density settlements (urban centres) and in close lower density areas. In this context, our results are in line with the urbanisation rate reported by the UN World Urbanisation Prospect (UN DESA, 2019), according to which the total urban population is 55% at the global level.

Fact 2: Both the proportion of metropolitan population and the average size of metropolitan areas tend to increase with income levels.

The share of people living in metropolitan areas increases with economic development. The average proportion of countries' population living in FUAs rises from 33% in low-income countries up to 70% in high-income countries. Such proportions are 50% and 55% for lower-middle and upper-middle countries, respectively (Table 3). This fact is clearly consistent with what we know from both theoretical and empirical literature, suggesting that urbanisation is a prominent feature of the spatial transformations generated from economic development (Desmet and Henderson, 2015). As countries develop, productivity gains in agriculture and the transformation of the economy with the shift towards manufacturing and service industries are important mechanisms driving rural–urban migration and the rise of metropolitan areas (Michaels et al., 2012).

Table 3
Population shares in FUAs, by FUA size and countries' income.

	<250K	250K-1M	1M-5M	>5M	Total
Low income	13%	7%	11%	2%	**33%**
Lower-middle	11%	12%	14%	13%	**5%**
Upper-middle	11%	14%	15%	15%	**55%**
High income	7%	17%	24%	22%	**7%**

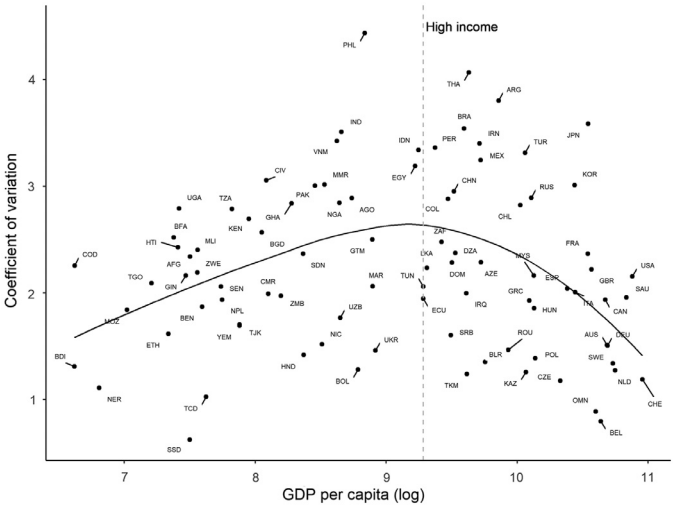

Fig. 8. Development level and the concentration of FUA population.
Note: 89 countries with more than 10 FUAs included. "High-income" according to the World Bank income classification.

Overall, the average size of metropolitan areas is larger in high-income countries. If we break down all FUAs in four size classes, each representing similar shares of population, we can clearly see that the proportions of population in the two largest classes – between 1 and 5 million and beyond 5 million inhabitants, respectively – tend to increase with income levels (Table 3). This evidence is coherent with the idea that city size increases in local productivity, human capital and travel speed (Duranton and Puga, 2019), all features which tend to increase with economic development.

Fact 3: At intermediate levels of development, a higher share of metropolitan population is concentrated in few large FUAs.

We observe a non-linear relationship between countries' average income per capita and the concentration of the FUA population across space. Fig. 8 plots such relationship, where the coefficient of variation measures the spatial concentration of the FUA population. For a

[16] Countries that are not recognised by the United Nations or whose recognition is controversial have been excluded, as well as countries with no GDP data available.

[17] We used the country groups as defined by the World Bank. On the method used to group countries by income, see: https://datahelpdesk.worldbank.org/knowledgebase/articles/378833-how-are-the-income-group-thresholds-determined (last access December 2019).

meaningful interpretation of the coefficient of variation, only countries with at least 10 FUAs entered the analysis. A simple linear regression confirms the inverse U-shaped relationship, even after controlling for total country population and using different measures of concentration of the metropolitan population.[18] According to such regression, the maximum metropolitan concentration occurs with levels of GDP per capita of about 10,000 USD.

This finding suggests that the dominance of a few large metropolitan areas over the remaining ones tends to increase from low to intermediate stages of development and then decreasing at higher income levels. The inverse U-shaped association between income per capita and metropolitan concentration seems to be driven by the existence of a few large FUAs dominating the metropolitan system rather than by a higher primacy – i.e. the proportion of total FUA population in the largest FUA. Our findings nuance previous evidence on the relationship between primacy and development. More specifically, our results confirm that primacy first increases and then decreases as income grows, coherently with the Shaks–Mera hypothesis (El-Shakhs, 1972).[19] However, the ratio between the population of the largest and that of the second-largest metropolitan areas does not show a bell-shaped relationship, more consistently with Lamelin and Polèse (1995).

Overall, our findings are coherent with the long-standing evidence that spatial transformations from economic development first increase polarisation and regional inequalities, followed by spread effects and a process of re-balancing spatial distribution of population and economic activities at more mature stages of development (Alonso, 1980; El-Shakhs, 1972).

Fact 4: Large shares of suburban population are a feature of rich countries.

The delineation of FUAs allows us to provide a consistent measure of the magnitude of suburbanisation at the global level. Overall, commuting zones surrounding urban centres represent 17% of the overall FUA population and 9% of the world population. From a geographical perspective, one third of FUA population lives in commuting zones in North America and Europe, while only 7% do so in African countries. The proportion of FUA population living in commuting zones reaches its peak in high-income countries (31%) and decreases to 18% and 10% in upper-middle and lower-middle income countries, respectively (Table 2; Fig. 9). In low-income countries, commuting zones represent less than 4% of FUA population. This gradient is robust to different specifications of the delineation model – as defined in Section 3.1 – and probability thresholds used to assign cells to FUAs. The same patterns hold even when adopting a constant probability threshold across world regions, with North America and Europe showing the highest shares of population in commuting zones.

The observed relationship between income levels and suburban population is consistent with the idea, originally formalised in the monocentric city model (Alonso, 1964; Muth, 1969; Kim, 2007), that richer households have flatter bid rent curves. Transportation might have a role in this respect. At higher development levels, a wider set of transportation choices becomes available and drives households with stronger preference for housing towards lower density areas outside the urban centre (LeRoy and Sonstelie, 1983).

Fact 5: Commuting zones are denser in poorer countries.

Functional urban areas in the developed world have a significant proportion of their population living in low-density (rural) settlements. To assess the distribution of FUA population in different types of settlement, we applied the 'Degree of Urbanisation' grid-cell classification

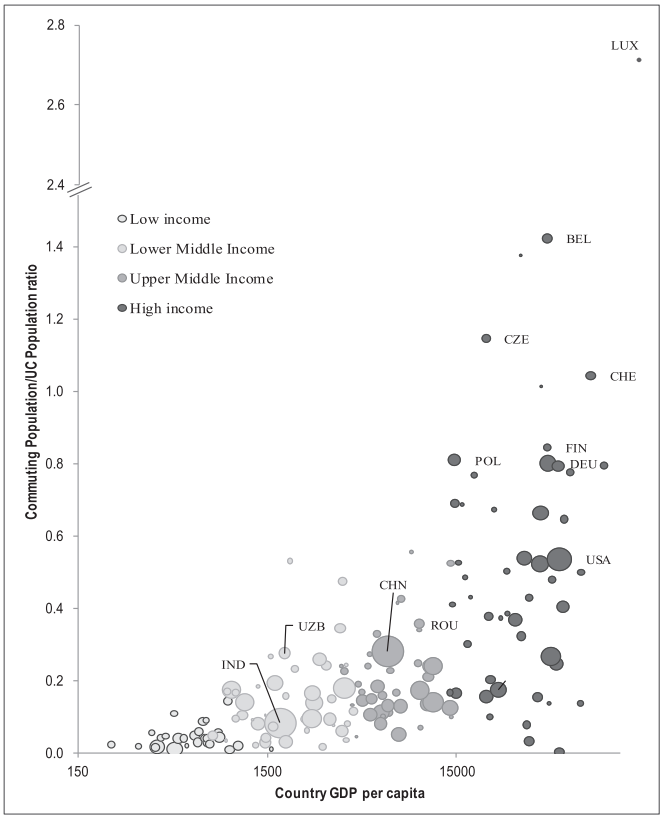

Fig. 9. Development level and importance of commuting zone, 2015.

Table 4
Proportion of FUA population by settlement density and income level.

	Urban centers (density > 1,500 inh./km^2) (%)	Towns & semi-dense areas (density > 300 inh./km^2) (%)	Rural (density < 300 inh./km^2) (%)
High income	69.4	21.8	8.8
Upper middle income	82.3	15.1	2.5
Lower middle income	89.6	9.3	1.1
Low income	96.4	3.1	0.5

(Dijkstra and Poelman, 2014). This classification reveals significantly higher settlement heterogeneity within FUAs in richer countries than in the rest of the world. In high-income countries, almost 22% of FUA population is located in towns and semi-dense areas (minimum density of 300 inhabitants per square kilometre), while about 9% is in rural settlements (density below 300 inhabitants per square kilometre) (Table 4). An almost negligible proportion of rural settlements characterises FUAs in lower middle and low-income countries.

Differences in the density of commuting zones across income levels are reflected by regional differences. Metropolitan areas in North America and Europe have the highest shares of population living in low density settlements (11% and 10%, respectively), followed by Oceania (6%). In all other world regions, the share of metropolitan population in low densities does not overcome 3%, on average, and is highest in Latin America. Possible explanations for the observed differences include that, at lower development levels, metropolitan areas require higher densities to ensure affordable services and transport mobility (Duranton and Turner, 2012).

4.2. Population dynamics and suburbanisation patterns

Fact 6: Population in FUAs grew faster than in other areas and, yet, one fifth of FUAs in the world shrank or stagnated.

[18] The R-squared of the regression for the coefficient of variation that includes the log of country GDP per capita and its quadratic value is 0.16. Results are consistent to alternative measures of concentration, such as the Gini coefficient of the FUA population, the Herfindahl-Hirschman Index, and the coefficients from the Zipf's law between rank and population for the largest 10 FUAs in the country. Results are available upon request.

[19] Regression results available upon request.

Between 2000 and 2015, the world FUA population increased by 21%, against 19% of growth in the remaining areas. The fastest growing FUAs are found in Africa, where FUA population increased by 46% between 2000 and 2015, against 38% in the whole continent.

Notwithstanding the general increase in population and the even stronger increase in the metropolitan population in practically all countries, more than one-fifth (22%) of FUAs experienced negative growth rates between 2000 and 2015. The patterns of declining metropolitan population mainly reflect regional demographic trends. The proportion metropolitan areas with declining population is highest in Europe (45%), followed by Asia (26%). In all other regions the share of declining metropolitan areas does not overcome 10%, with the lowest proportion observed in Oceania (3.6%) and Latin America (5%).

Population decline in metropolitan areas can also reflect a redistribution of population across different metropolitan areas within the same country. In China, for example, the migration of people from the West to the East sheds some light on why population in 42% of Chinese FUAs declined in the considered period.

Measuring population dynamics using FUAs rather than just high-density urban centres proofs particularly useful to distinguish actual urban shrinkage from a more common process of suburbanisation and redistribution of the metropolitan population from the urban centre to its commuting shed. If we had measured city decline only looking at urban centres, we would observe that the proportion of declining cities is 28% rather than 22% of total FUAs worldwide.

Fact 7: Globally, larger FUAs have grown faster.

Between 2000 and 2015, population growth occurred at a faster pace in larger metropolitan areas. FUAs with over one million inhabitants grew 0.16 percentage points faster than national population growth and a half of a percentage point faster than FUAs with less than one million inhabitants. FUAs with over five million people grew even faster, one percentage point above smaller FUAs. Taken together, the 94 FUAs with more than five million inhabitants added more than 224.4 million people in the period, more than doubling the 100.5 million in 6,092 FUAs with less than 250,000 inhabitants.

Differences in growth rates explain the observed changes in the proportion of the world population in metropolitan areas of different sizes. The share of world population in FUAs of over five million inhabitants increased in the fifteen years considered by almost one percentage point (Fig. 10), from 13% to 14%. While FUAs with less than one million inhabitants experienced on average a positive change in population, on average, their share over total FUA population slightly decreased in the observed period. Overall, this pattern suggests that a process of concentration of population toward the largest metropolitan areas has occurred worldwide in the last fifteen years.

Fact 8: Suburban population experienced a widespread increase and offset urban centre decline in 17% of FUAs.

The concentration of population in large metropolitan areas was coupled with a slow but worldwide-spread increase of suburban population. As Table 2 shows, between 2000 and 2015, the share of population in commuting zones increased in most countries, consistently with previous evidence for developed countries (Veneri, 2018). The proportion of population in commuting zones over total FUA population increased by 1.6 percentage points in the period, on average. In relative terms, that shift towards commuting zone was highest in Asia (1.9 p.p.). These figures are most conservative, as we keep boundaries of both FUAs and urban centres constant at 2015, thus using larger units than those that were likely to exist at the beginning of the period. Possible explanations for the observed decentralisation of metropolitan population can be related to improved intra-metropolitan mobility, such as through availability of new roads (Duranton and Turner, 2012).

Distinguishing growth in urban centres and commuting zones makes it possible to identify different patterns of metropolitan growth at the global scale. Almost three quarters of FUAs experienced growth in both urban centres and in their respective commuting zones (Table 5). Within this group, population in commuting zones grew at higher speed than in urban centres in 88% of FUAs, although in absolute terms commuting zones represented a smaller population. About 17% of FUAs worldwide experienced a decline in the urban centre and an increase in the commuting zone population. For 68% of FUAs in this group, growth in commuting zones more than compensated population shrinkage in urban centres. On the other hand, 10% of FUAs had population decline in both urban centres and commuting zone. Finally, only in a negligible number of FUAs (1%) commuting zones declined while urban centres grew. Such a small group suggests that metropolitan centralisation has not been a common pattern since the turn of the millennium, confirming previous projections (Cheshire, 1999).

Table 5
Population growth in urban centres vs. commuting zones, 2000–15.

FUAs' Population growth pattern (2000–15)	*Share of FUAs (%)*	*Share of total FUA population (%)*	*Median FUA yearly growth rate (%)*
Both urban centres and commuting zone grow	72	83.9	1.3
Both urban centres and commuting zone decline	10	6.1	−0.7
Urban centres grow while commuting zone declines	1	1.1	0.3
Urban centres decline while commuting zone grows	17	8.7	−0.2

Note: FUAs in which at least one urban centre was not existing in 2000 were not included.

Table 6
Decomposition of population growth of FUAs, 2000–15.

Income group	Densification	Growth in urban centres' expansion area	Growth in commuting zones
Low Income	83%	13%	4%
Lower Middle	73%	13%	14%
Upper Middle	68%	13%	19%
High Income	56%	9%	35%
World	70%	13%	17%

Fact 9: Globally, densification of existing urban centres accounted for half of total FUA population growth.

Zooming into dynamics of population within FUAs, our data allow identifying three different components of metropolitan growth, which, to our knowledge, have not yet been documented in a systematic way. The first component – 'densification' – accounts for growth within the urban centres' boundaries, where the latter are defined at the beginning of the period (2000). The second component accounts for growth in the area of expansion of urban centres between the two points in time. In other words, such a component takes into account the areas that were previously of lower to medium density and that subsequently became denser and part of a new, larger urban centre. Finally, the third component accounts for the growth within the commuting zone, as defined at the end of the period (2015). As such, the third component can be looked at as a conservative measure of suburbanisation.

Overall, while population has been slowly shifting towards commuting zones, densification alone accounted for 70% of growth of metropolitan population. Combined with the previous fact, this suggests that metropolitan areas are becoming denser on average, both in urban centres and commuting zones. Table 6 shows the decomposition of FUAs' population growth for the entire world and by level of income for 2000–15. As expected, densification is highest in low-income countries (83%) and lowest in high-income countries (56%), while a reverse gradient is observed for the contribution of commuting zones. On the other hand, the growth in urban centres' expansion was remarkably consistent across countries, accounting for 9% in high-income countries, and for 13% in the rest of the world. From a regional point of view, the contribution of commuting zones to metropolitan growth was highest in North America (41%), followed by Europe and Central Asia (36%). On the other hand, in Sub-Saharian Africa and Middle-East/North Africa, commuting zones contributed to 7% and 13% of total metropolitan growth, on average.

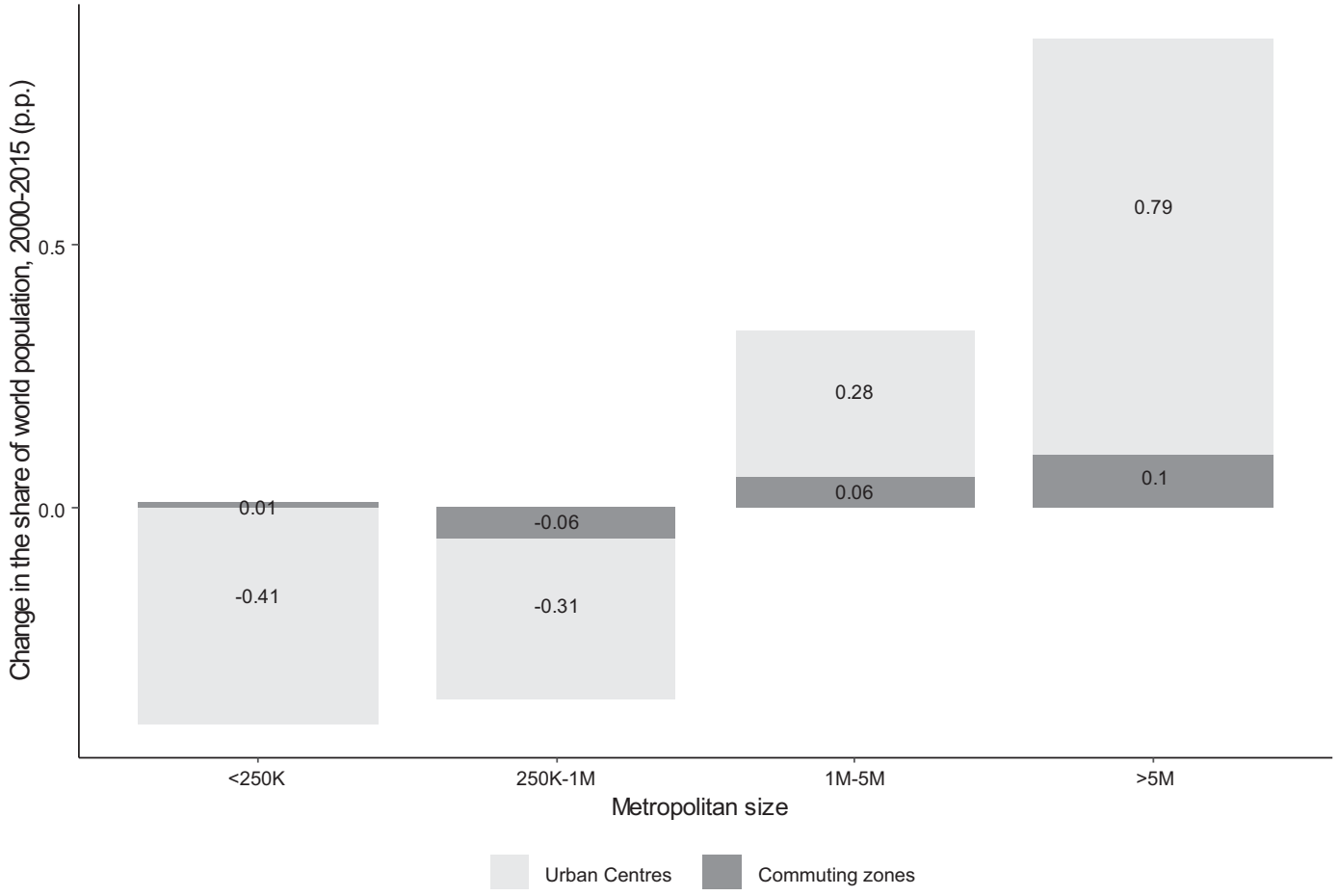

Fig. 10. Change in the proportion of world population in metropolitan areas (2000–15), by metropolitan size.

5. Concluding remarks

Any assessment of global trends in urban development needs to rely on a consistent definition of what is a city and its economic area of influence. This paper presented a novel method to define commuting zones around urban centres at the global scale in the absence of commuting flows data. The method relies on the concept of FUA as defined by the OECD and the European Union (OECD, 2012; Dijkstra et al., 2019).

By using global grids of population and travel impedance at one-km^2, the method uses the information available on the commuting zones in 31 OECD countries to predict the extent of commuting zones all around the world. The estimated extent of FUA boundaries in OECD countries turned out to be satisfactorily accurate in terms of population with respect to the original FUAs identified by aggregating local administrative units to the high-density urban centres based on the intensity of commuting flows.

One major feature of the proposed method is that it is 'people-based', given that FUAs are defined as agglomerations of people. Such a definition has the advantage of being more independent with respect to differences in development levels when performing a global analysis. Other approaches relying on aspects related to human activities, such as built-up areas or lights, can be more sensitive to the context under analysis.

The method proposed in this paper relies on a training set that is built on a sample of countries which are mostly developed. This might represent a limitation in the capacity to predict with sufficient accuracy the extent of FUAs in the least developed countries. However, the global impedance grid used to compute the travel time of each cell to its closest high-density urban centre embeds the different costs of moving from different points in space according to the level of infrastructure and the geographical characteristics of the specific area under study. In addition, calibrating the model by world region and using country-level GDP per capita can help mitigate a possible bias in the country sample used for the training set.

The delineation of FUAs across the whole globe made it possible to assess some key features of metropolitan areas and to look at population dynamics in urban centres and commuting zone separately. According to our estimations, FUAs represent about 53% of the world population, out of which 17% are located in the commuting sheds of urban centres. Measuring population growth using FUAs makes it possible to capture the actual growth of the 'economic' city and avoid confusing actual population decline with a decentralisation of population towards commuting zones.

Our estimations confirm that population in metropolitan areas have grown faster than elsewhere in the last fifteen years, on average, and that larger metropolitan areas have grown faster than smaller ones. Growth of population was stronger in the commuting zones than in dense urban centres, on average, suggesting that metropolitan growth has occurred in parallel to a process of decentralisation of the urban centres' population in most parts of the world. Identifying the factors behind the observed global heterogeneous patterns of growth between and within metropolitan areas will continue to offer relevant questions for future research. The global, consistent definition of metropolitan areas provided with this paper might be helpful to further address some of those questions.

CRediT authorship contribution statement

Ana I. Moreno-Monroy: Methodology, Software, Validation, Formal analysis, Writing - original draft. **Marcello Schiavina:** Methodology, Data curation, Formal analysis, Software. **Paolo Veneri:** Conceptualization, Formal analysis, Writing - original draft, Supervision, Project administration.

Acknowledgments

This work benefited from financial support by the European Commission, Directorate-General for Regional and Urban Policy. We are grateful to the editor, Gilles Duranton, two anonymous referees, as well as Bernardo Alves, Lewis Dijkstra, Dante Donati, Vernon Henderson, Rafael Pereira, Toni Venables, conference participants, and OECD and JRC colleagues for their valuable input and comments. Special thanks to Luca Maffenini for excellent implementation support. We are indebted to Maarten Bosker, Jane Park, and Mark Roberts for kindly sharing their data on Indonesia. Every error remains the authors' responsability. The views expressed herein are those of the authors and do not necessarily reflect the views of the OECD, the European Union or of their respective member countries.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.jue.2020.103242.

Annex A. Data sources and description

Population density and human settlements

For the definition of population density and human settlements at the global level we rely on the Global Human Settlement Population (GHS-POP, Schiavina et al., 2019) and settlement model (GHS-SMOD, Pesaresi et al., 2019) containing information on population by one-km^2 grid cells and their types in equal area Mollweide projection (EPSG: 54,009). The data are provided by the Joint Research Centre of the European Commission and are available for download at https://ghsl.jrc.ec.europa.eu/datasets.php. Technical details can be found in Florczyk et al. (2019).

We choose to use GHS-POP layer (Schiavina et al., 2019) for population distribution as it is a global and harmonized layer using a well-established dasymetric disaggregation procedure (Freire et al., 2016) based on open global datasets:

- the GPWv4.10 (CIESIN 2017), a global harmonized census dataset provided by the centre for International Earth Science Information Network (CIESIN - Columbia University) that is aligned with UN Population data at country level, as input population values;
- the latest release of GHS-BUILT (Corbane et al., 2018; 2019) a global harmonized layer of built-up surface, as target for population distribution.

As with any global product, such datasets are not exempt from issues due to the accuracy of the target built-up surface layer and quality of census data for many countries, but methodologies are continuously improved to reduce errors (Corbane et al., 2019) and with procedures to detect and mitigate major discrepancies and anomalies occurring in geospatial input population data (Freire et al., 2018).

Travel time calculations

The impedance matrix has global extent (30 arc-second resolution, in WGS84, EPSG: 4326, coordinate system), southward limited up to the 60 °S parallel (i.e. excluding Antarctica, where there are no relevant human settlements). Its coordinate system is in WGS84 (EPSG: 4326) with a 30 arc-seconds resolution. Cell values represent the inverse of the estimated speed to go through each cell. To handle the different projections in the distance calculation, the relevant population cells and settlement edges cells are converted into points (using the cell centroids). The points' Mollweide coordinates are then unprojected to WGS84 geographical coordinates. Every WGS84-cell of the impedance layer in which the unprojected centroids fall inside corresponds to a population cell and settlement edge cell. The grid is freely available for download at https://map.ox.ac.uk.

Distances, expressed as travel time, are computed for all 8-connectivity paths (which assumes it is possible to travel from a cell A to all 8 cells surrounding A from each relevant population WGS84-cell to all settlement edges WGS84-cells B, selecting the shortest path following the Dijkstra algorithm (Cormen et al., 2001). Each path from A to B is described by a sequence of WGS84-cells on the impedance matrix and the distance is computed as the sum of time requested by each transition from each WGS84-cell to the adjacent one. Each transition time is obtained as the arc-length (i.e. approximating the Earth as a sphere) between the two WGS84-cell centroids divided by the average speed of the two WGS84-cells.

To calculate travel times, we start by clipping the global travel impedance matrix on the country extent and calculating the travel time (in minutes) between each cell and the edge of all urban centres in each country. Calculating the travel time to the edge of urban centres considerably increases the computational weight compared to calculating distances to a single point in each FUA (e.g. the cell with highest density) but was preferred for two reasons. First, urban centre centroids might fall by chance in remote areas (e.g. the top of the Corcovado mountain in Rio de Janeiro) and could thus artificially inflate travel times and inducing the selection of a less optimal destination. Second, the same population cell could be far away from the centroid, if the adjacent urban centre is very large, and closer to the centroid of another small but distant urban centre.

A note on the reliability of this source: the impedance layer provide the average speed needed to cross a given cell and this is just an approximation of reality as, given a square kilometre the speed it is possible to reach crossing it could greatly vary according to the real path and the direction. Other possible cases are places characterized by the presence of barriers (slopes, water, etc.) in parts of a one-km^2 cell. In the impedance layer, these barriers will affect the whole km^2, thus reducing the speed of crossing, reaching or leaving it even if in reality there are paths not influenced by the mentioned barriers.

Other sources

Baseline FUA boundaries: The baseline FUA boundaries for 31 countries obtained following the method outlined in OECD (2012) and Dijkstra et al. (2019) can be found at https://www.oecd.org/cfe/regional-policy/functionalurbanareasbycountry.htm.

Country boundaries: As country boundaries we use the GADM v2.8 country boundaries available at https://gadm.org/.

Country-level additional information: Vehicles per capita data from https://en.wikipedia.org/wiki/List_of_countries_by_vehicles_per_capita; GDP per capita was downloaded from the World Bank Open Data website, https://data.worldbank.org/indicator/NY.GDP.PCAP.CD. Gaps in GDP data were filled for the following countries (years/source): Bermuda (2013), Channel Islands (2007), Curacao (2011), Cayman Islands (average 2005–2006), Eritrea (2011), New Caledonia (2000), Korea, Dem. People's Rep. (2016, Wiki), French Polynesia (2000), Syrian Arab Republic (2000), Venezuela, RB (2014), Western Sahara (CIA Factbook 2007).

Nighttime lights: For the sum of nightlights in urban centres the source is Version 1 Nighttime VIIRS (Visible Infrared Imaging Radiometer Suite) Day/Night Band Composites suite produced by The Earth Observations Group (EOG) at NOAA/NCEI, available at https://www.ngdc.noaa.gov/eog/viirs/download_dnb_composites.html#NTL_2015. These grids span the globe from 75 N latitude to 65S and have a resolution of 15 arc-second in WGS84 geographic coordinates (EPSG 4326). The yearly "vcm-orm-ntl" (VIIRS Cloud mask - Outlier Removed - Nighttime Lights) layer was selected, showing the cloud-free average radiance emitted by Earth (expressed as nW cm^{-2} sr^{-1}) with outlier removal process to filter out fires and other ephemeral lights. This layer has been warped to the GHS-SMOD grid by oversampling at 50 m in Mollweide projection (EPSG 54009), with nearest neighbor method, then aggregated at one-km^2 by averaging values.

Annex B. Model selection

Table B.1
Summary statistics.

	Mean	Standard deviation	Median	Maximum	Minimum
Travel time (minutes)	34	111	17	8,755	0
Area urban centre (km^2)	343	778	84	5,633	5
Population urban centre (persons)	1,050,271	3,315,001	197,761	33,028,731	50,070
Nighttime lights urban centre (light intensity)	29	14	29	118	4
Cell population (persons)	877	832	617	41,202	300
GDP per capita (USD, PPP)	38,893	16,863	41,324	101,447	6,085
Cars per 1000 inh.	609	158	586	797	148

Table B.2
Stepwise results (extract).

	BIC	Maximum clustered p-value (among considered regressors)	Rank	Δ BIC from previous step	
Step 1					
dist	*366335*	*0.000*	***	*1*	–
size_uc_area	605844	0.000	***	2	–
size_uc_size_cell_pop	614437	0.000	***	3	–
size_uc_nl	644640	0.908		7	–
size_cell_pop	644551	0.020	*	6	–
GDP	630546	0.000	***	4	–
cars	635323	0.000	***	5	–
...					
Step 5					
$dist^3$ size_uc_area size_cell_pop	339933	0.424		6	−59.9
$dist^2$ $size\_uc\_area^2$ size_cell_pop	339952	0.615		7	−41.7
$dist^2$ size_uc_area size_cell_pop2	339888	0.000	***	5	−105.6
$dist^2$ size_uc_area size_cell_pop size_uc_size_cell_pop	339342	0.416		3	−651.1
$dist^2$ size_uc_area size_cell_pop size_uc_nl	339258	0.022	*	2	−735.6
$dist^2$ size_uc_area size_cell_pop GDP	*338956*	*0.001*	**	*1*	*−1036.8*
$dist^2$ size_uc_area size_cell_pop cars	339747	0.094		4	−245.8
Step 6					
$dist^3$ size_uc_area size_cell_pop GDP	338909	0.461		5	−47.3
$dist^2$ $size\_uc\_area^2$ size_cell_pop GDP	338948	0.746		7	−8.6
$dist^2$ size_uc_area size_cell_pop2 GDP	*338906*	*0.002*	**	*4*	*−50*
$dist^2$ size_uc_area size_cell_pop GDP size_uc_size_cell_pop	338925	0.581		6	−31.6
$dist^2$ size_uc_area size_cell_pop GDP size_uc_nl	338047	0.087		2	−909.3
$dist^2$ size_uc_area size_cell_pop GDP^2	*332719*	*0.007*	**	*1*	*−6236.7*
$dist^2$ size_uc_area size_cell_pop GDP cars	338746	0.111		3	−210.2
Step 7					
$dist^3$ size_uc_area size_cell_pop2 GDP	338858	0.453		2	−48.9
$dist^2$ $size\_uc\_area^2$ size_cell_pop2 GDP	338899	0.753		4	−7.5
$dist^2$ size_uc_area size_cell_pop3 GDP	338919	0.861		5	13
$dist^2$ size_uc_area size_cell_pop2 GDP size_uc_size_cell_pop	338880	0.604		3	−26.4
$dist^2$ size_uc_area size_cell_pop2 GDP cars	338687	0.105		1	−219
Step 8					
$dist^2$ size_uc_area size_cell_pop2 GDP dist:size_uc_area	338819	0.174		4	−87.6
$dist^2$ size_uc_area size_cell_pop2 GDP dist:size_cell_pop	338908	0.271		6	1.1
$dist^2$ size_uc_area size_cell_pop2 GDP dist:GDP	338559	0.839		2	−347.9
$dist^2$ size_uc_area size_cell_pop2 GDP size_uc_area:size_cell_pop	338797	0.790		3	−109.3
$dist^2$ size_uc_area size_cell_pop2 GDP size_uc_area:GDP	338435	0.593		1	−471.1
$dist^2$ size_uc_area size_cell_pop2 GDP size_cell_pop:GDP	338894	0.269		5	−12.4

Note: dist = travel time; size_uc_area = area of urban centre; size_cell_pop = cell population; GDP = GDP per capita. For each step, the selected model is shown in italics. Steps 2–4 are omitted for presentation purposes but are available upon request. The final selected model is indicated in bold text. Step 6 shows the discarded model with second-degree polynomial for GDP. Urban-centre-size proxy variables representing its area, population and nightlights are size_area; size_pop and size_nl respectively. Exponent of variables show the degree of the polynomial considered.

Table B.3
Regression results.

	Estimate	Std. error	z value	Pr(>|z|)
Travel time	−0.442	0.120	−3.673	0.000
Travel time^2	−0.301	0.029	−10.422	0.000
Area urban centre	0.442	0.048	9.144	0.000
Cell population	0.714	0.242	2.945	0.003
Cell population^2	−0.070	0.018	−3.829	0.000
GDP country	0.263	0.085	3.106	0.002

Note: Estimation based on a Generalized Linear Model with errors clustered by urban centre. Intercept included but not shown. All shown variables are log-transformed. Estimates corresponding to raw polynomials are shown for ease of interpretation, but estimates used in the implementation are obtained using polynomials orthogonal to the constant polynomial of degree 0 to decrease multicollinearity bias.

Annex C. Comparison with a nested specification

The data used to estimate Eq. (2) has a nested or block structure, since each cell is assigned to an urban centre and urban centres belong to countries. In this case, the independence assumption across cells may not hold. This annex compares the performance of the baseline regression with respect to an alternative specification that takes into account the nested structure of the data is compared with a mixed-effects logistic regression model with random intercepts in two levels, so that each urban centre has its own random intercept varying within each country. The estimation is done via a Maximum Likelihood estimator using the *glmer* function of the *lme4* package in R (Bates et al., 2015). Table C.1 shows the regression results for the fixed effects component of the model.

Table C.1
Fixed-effects estimation results of mixed effect logistic model with random intercepts by urban centre and country.

	Estimate	Std. error	z value	Pr(>|z|)
Distance	1.970	0.146	13.475	<2e-16
Distance^2	−0.288	0.025	−11.726	<2e-16
Area urban centre	−0.591	0.006	−106.688	<2e-16
Cell population	0.359	0.062	5.812	6.18e-09

Note: Estimation based on a Generalized Linear Model with errors clustered by urban centre. Number of obs: 466,361, groups level 1 (urban centre): 1,287; level 2 (country): 31. Estimates corresponding to raw polynomials are shown for ease of interpretation, but estimates used in the implementation are estimated using polynomials orthogonal to the constant polynomial of degree 0 to decrease multicollinearity bias.

Table C.2
Performance comparison between baseline model and nested model.

Performance metric	Country (n = 31)			Urban centre (n = 1,287)		
	Median nested	Median baseline	p-value	Median nested	Median baseline	p-value
Sensitivity	0.920	0.930	0.316	0.901	0.950	0.000
Specificity	0.839	0.793	0.053	0.900	0.840	0.000
Balanced Accuracy	0.847	0.834	0.600	0.832	0.815	0.624

Table C.2 shows the median results of comparing three performance statistics for the baseline and nested specifications across the 31 countries and 1,287 urban centres in the baseline sample. These metrics are based on cells inside and outside FUAs as predicted by the model and not as actually observed after drawing the boundaries, so they may differ slightly from the performance statistics in Table 1. The p-value of a two-sided *t*-test of the difference between the two alternatives is included for reference.

According to these results, at the country level both models perform equally well as the median performance across 31 countries is not statistically different. At the urban centre level, the comparison results show that while the baseline specification performs better in terms of sensitivity, the nested specification performs better in terms of specificity.

Fig. C.1 shows the results by country and urban centre for the nested and baseline specifications. At the country level, specificity is one percentage point higher or more in Greece, South Korea, Italy, Hungary, Portugal, Colombia and Slovenia when using the nested model instead of the baseline one, whereas sensitivity is one p.p. higher or more in Chile, South Korea, Netherlands and Japan. These results show that not taking into account random intercepts negatively affects performance in terms of specificity (the correct identification of actual negatives), but taking them into account negatively affects sensitivity (the correct identification of actual positives). The countries for which specificity is lower do not share a salient characteristic (e.g. size, geographical position) that would give clues on which variables could be added to the baseline specification. In fact, performance tests based on urban centre characteristics such as size do not reveal any salient pattern (results available upon request).

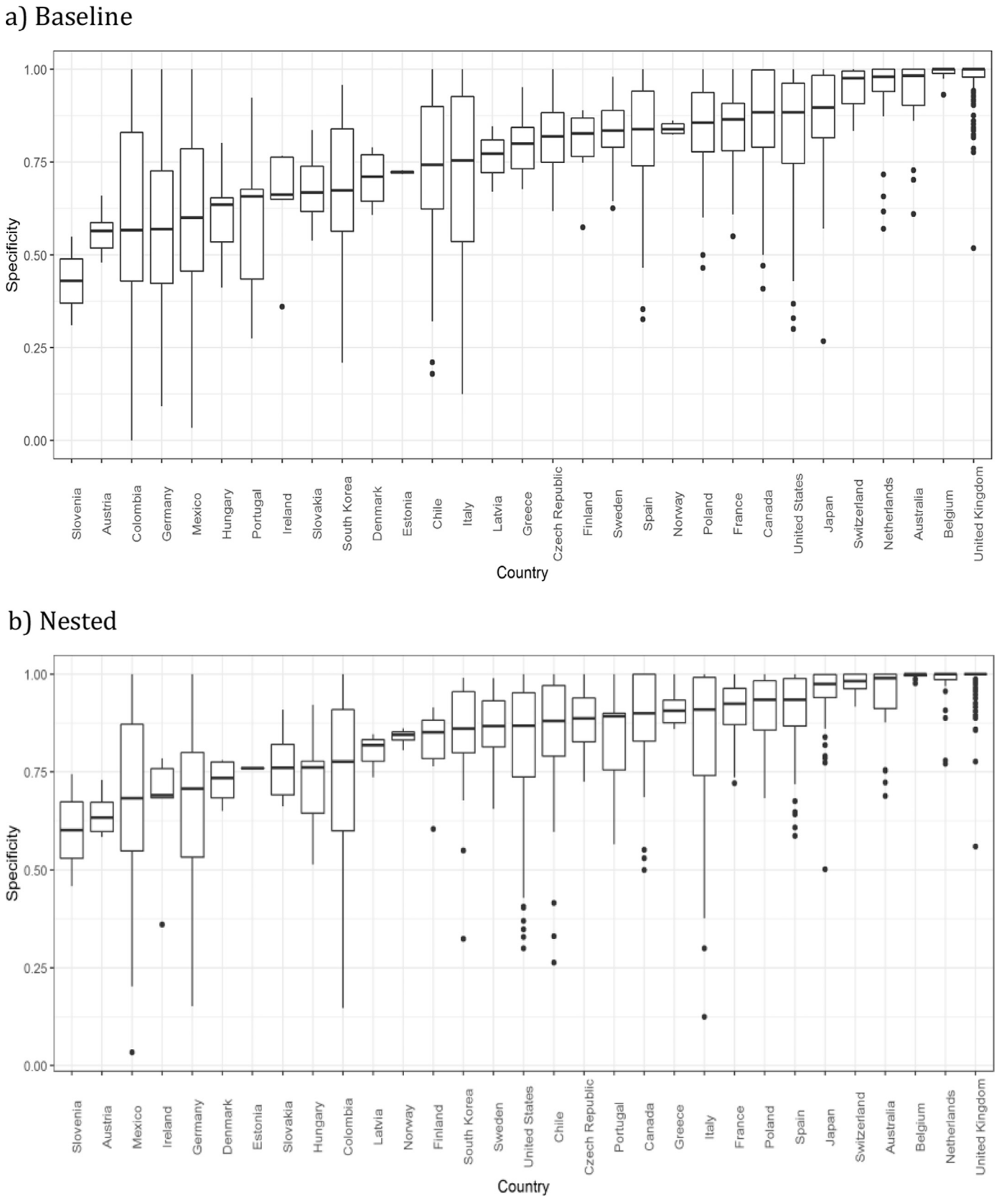

Fig. C.1. Baseline versus nested specification specificity results by country and FUA.

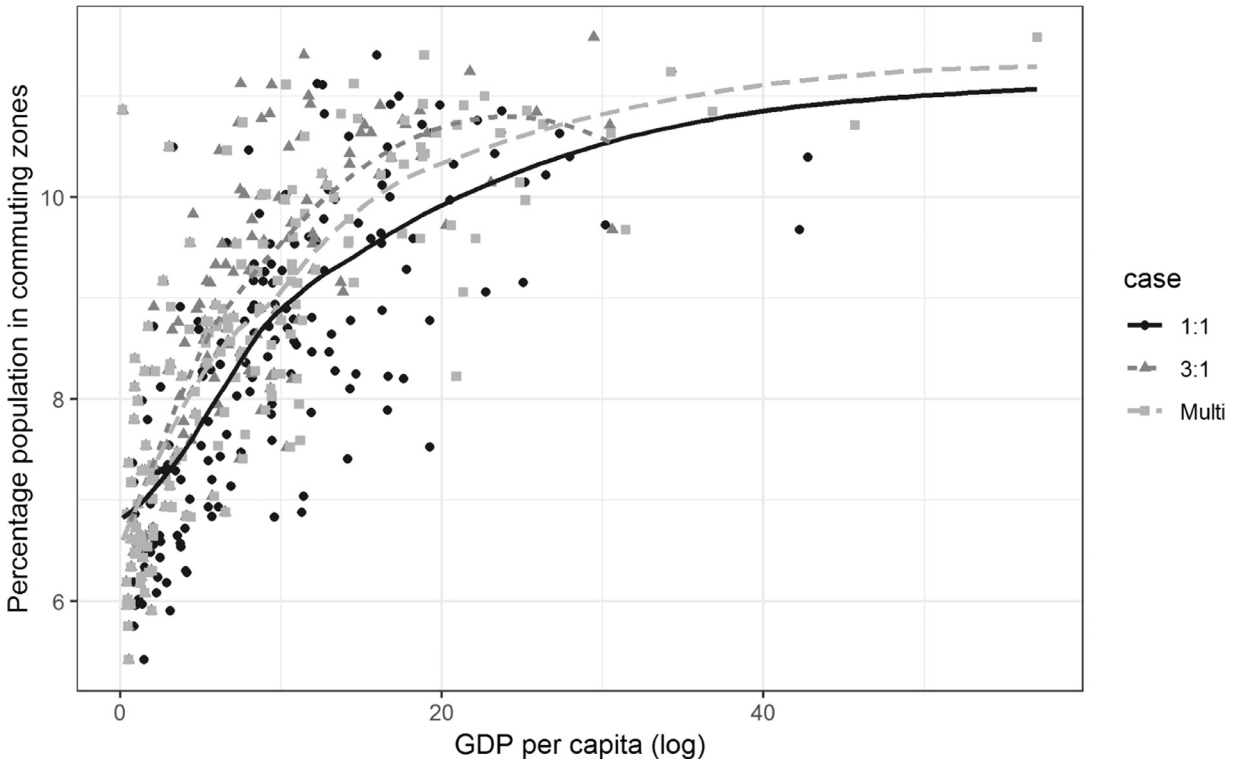

Fig. D.1. Percentage of people in commuting zones versus GDP per capita by version.

Annex D. Optimal threshold determination

This annex compares global boundaries obtained using different optimal probability thresholds and positive versus negative error weights using the same data and econometric specification.

The following three options for positive versus negative weight were explored: including weighting false negative and false positive errors equally (1:1) or twice/three times as much as false negative error (2:1 or 3:1). Using the full sample information leads to an optimal calibrated threshold of 0.53 when false negative and false positive errors are weighted equally (version 1:1). Weighting false positive error twice as much as false negative error (version 2:1) leads to a calibrated threshold of 0.66. Weighting false positive error three times as much as false negative error leads to a calibrated threshold of 0.75 (version 3:1).

To obtain thresholds by world region, we calibrated optimal probability thresholds by clustering countries with baseline FUAs across UN world regions (version "multi"). Table D.1 summarizes the UN world regions with their corresponding thresholds and the population values used in each case.

Table D.2 summarizes the performance results for the four versions. Consider versions 1:1, 2:1 and 3:1 for the moment. Version 1:1 would be chosen based on sensitivity and false negative error, and version 3:1 would be chosen based on specificity and false positive error.[20] These results make sense given that a lower probability value will allow a higher likelihood of identifying cells inside FUAs correctly at the expense of more false negative that yield boundaries that are "too large". On the contrary, version 3:1 is less likely to correctly identify cells outside FUAs but more likely to yield boundaries that are "too small" because a higher probability threshold on cells inside FUAs.

These performance results can be compared with version "multi". As can be expected from the combination of lower and higher probability

[20] As the performance of version 2:1 lies in between versions 1:1 and 3:1, we will omit it henceforth.

Table D.1
Calibrated thresholds by world region.[27]

	Regional optimal threshold	Population used in calibration	Total 2015 population (source: UN WPP 2017)
Northern America	0.45	272,431,892	357,700,770
Eastern Asia	0.74	151,957,225	1,612,287,066
Eastern Africa	*0.74*	–	394,477,342
Middle Africa	*0.74*	–	151,951,734
Northern Africa	*0.74*	–	223,891,511
Southern Africa	*0.74*	–	62,633,712
Western Africa	*0.74*	–	353,223,876
Caribbean	*0.74*	–	43,017,234
Central America	0.42	104,580,109	172,908,048
South-Eastern Asia	*0.74*	–	633,497,753
South-Central Asia	*0.74*	–	1,890,288,217
Western Asia	*0.74*	–	257,230,985
Eastern Europe	0.42	49,493,239	292,942,786
Northern Europe	0.41	75,813,442	102,357,768
Southern Europe	0.64	110,016,518	152,349,077
Western Europe	0.53	152,080,370	190,792,170
Australia/New Zealand	0.41	20,716,438	28,497,494
Melanesia	*0.74*	–	9622,827
South America	0.60	58,103,162	418,447,713
Micronesia	*0.74*	–	526,344
Polynesia	*0.74*	–	684,616

Table D.2
Performance results across threshold choices.

Performance metric	Version 1:1	Version 2:1	Version 3:1	Version multi
Sensitivity	0.834939	0.750085	0.673099	0.83964
Specificity	0.853254	0.906833	0.935251	0.85842
Balanced Accuracy	0.844096	0.828459	0.804175	0.84903
False Positive Error	0.146746	0.093167	0.064749	0.14158
False Negative Error	0.165061	0.249915	0.326901	0.16036

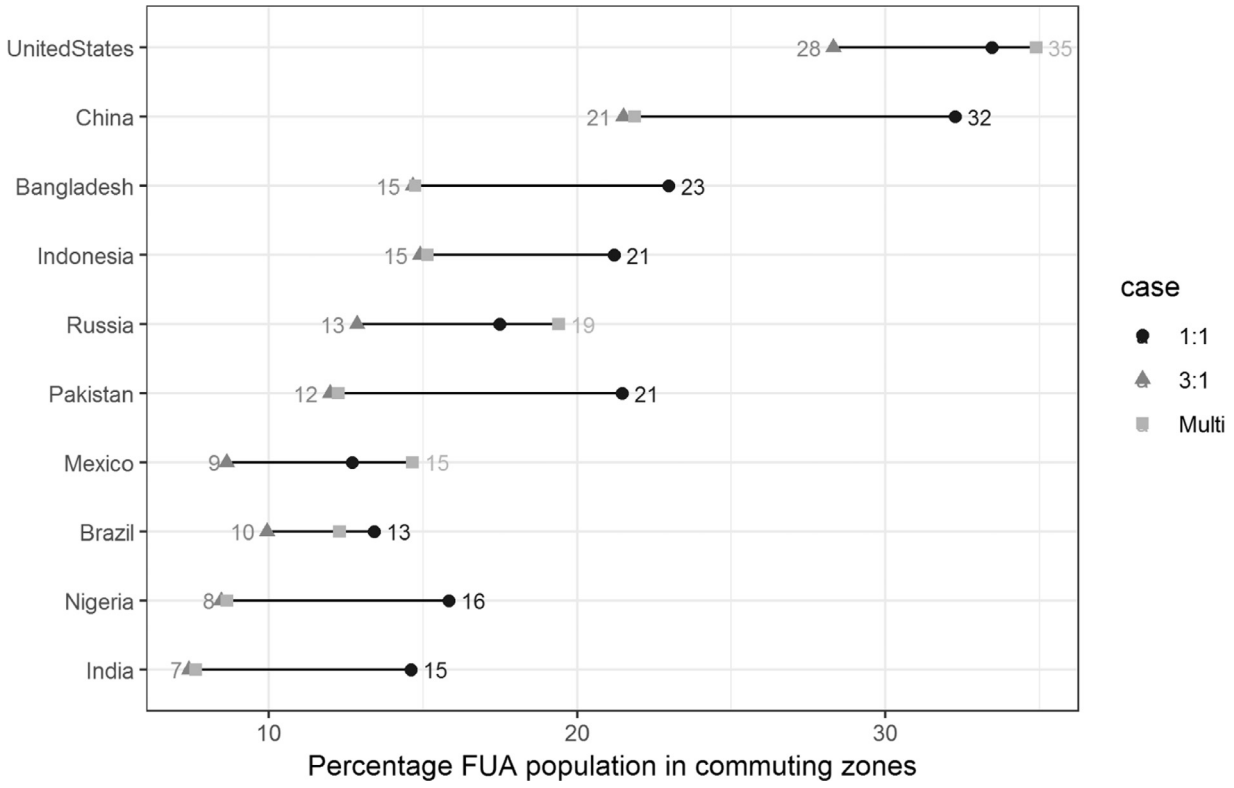

Fig. D.2. Share of FUA population in commuting zones by version, 10 most populated countries.

thresholds, the performance metrics of version "multi" compared to version lower (higher) in specificity and higher (lower) in false positive error compared to version 3:1 (1:1). Version "multi" has similar – and even better – performance metrics than version 1:1 in terms of sensitivity and false negative error, and outperforms all options with respect to balanced accuracy. Even if the average performance metrics are similar, version 1:1 may be more likely to over-estimate FUAs outside baseline countries, especially in radically different contexts to baseline countries such as Africa and South-East Asia. On the other hand, version 3:1 may be likely to predict smaller FUAs outside baseline regions and also may under-estimate boundaries in benchmark countries because those have a smaller optimal threshold in version "multi".

The comparison of summary statistics aggregated by region and level of development across all models gives additional insight regarding the effect of optimal thresholds. Version 1:1 predicts more people in FUAs than version 3:1 mainly due to many more people in commuting zones, with an absolute difference of 346,476,719 people worldwide, out of which: 70% are in Asia; 13% in Europe; 7% in Africa; and the rest in other regions of the world. By levels of development, the corresponding percentages are 76% in less developed countries, 20% in developed countries and 3% in least developed countries.

Compared to version 1:1 and as it could be expected from the calibrated threshold values, version "multi" gives similar predictions for the share of people in commuting zones for Europe (−0.04 p.p. difference), slightly larger shares for Latin America & Caribbean (0.88), North America (1.43) and Oceania (1.67), and smaller shares for Africa (−6.02) and Asia (−8.02). Compared to version 3:1, which sets a relatively high probability threshold for all countries, version "multi" predicts larger percentages for Europe (8.78), Latin America & Caribbean (3.6), North America (6.52), Oceania (5.77), and similar percentages for Africa (0.21) and Asia (0.27).

Version 1:1 predicts larger percentages of people in commuting zones with respect to total population than version 3:1 throughout the GDP per capita distribution. This shows the relationship between GDP per capita and percentage of population in commuting zones. Version 3:1 predicts much lower commuting population shares for the richest countries than versions 1:1 and "multi". This means that version 3:1 requires a higher level of development to reach the same percentage of commuting than version 1:1, but relatively less so after relatively high percentages of commuting population (∼ >20%). Version "multi" on the other hand displays a similar behaviour than version 3:1 at bottom and middle levels of income per capita, while still predicting relatively large percentages at the top of the distribution as version 1:1.

Fig. D.1 compares the share of FUA population in commuting zones in top 10 most populated countries of the world shows that version 1:1 has larger commuting percentages compared to version 3:1 and "multi". These differences can be quite significant for large countries with relatively low levels of income, corroborating that version 1:1 likely over-estimates the number of people in commuting zones outside high-income countries (Fig. D2).

Annex E. Robustness to merging rules

Change in population threshold at which two FUAs are not merged

The merging procedure based on distance between urban centres is not applied whenever the population of either urban centre is at least 500 thousand people. To test the effect of changing this population threshold two options are considered: 1) not imposing any population threshold; 2) increasing the threshold to 1 million people. In both cases the merging distance is kept constant at 5 km (baseline case).

As in Section 3.4, the rank-size rule is used to identify cases in which the rule significantly affect the results in terms of city size distribution. Fig. E.1 shows the results for cases 1) and 2) against the baseline case and highlights cases where the difference in slope coefficients is larger

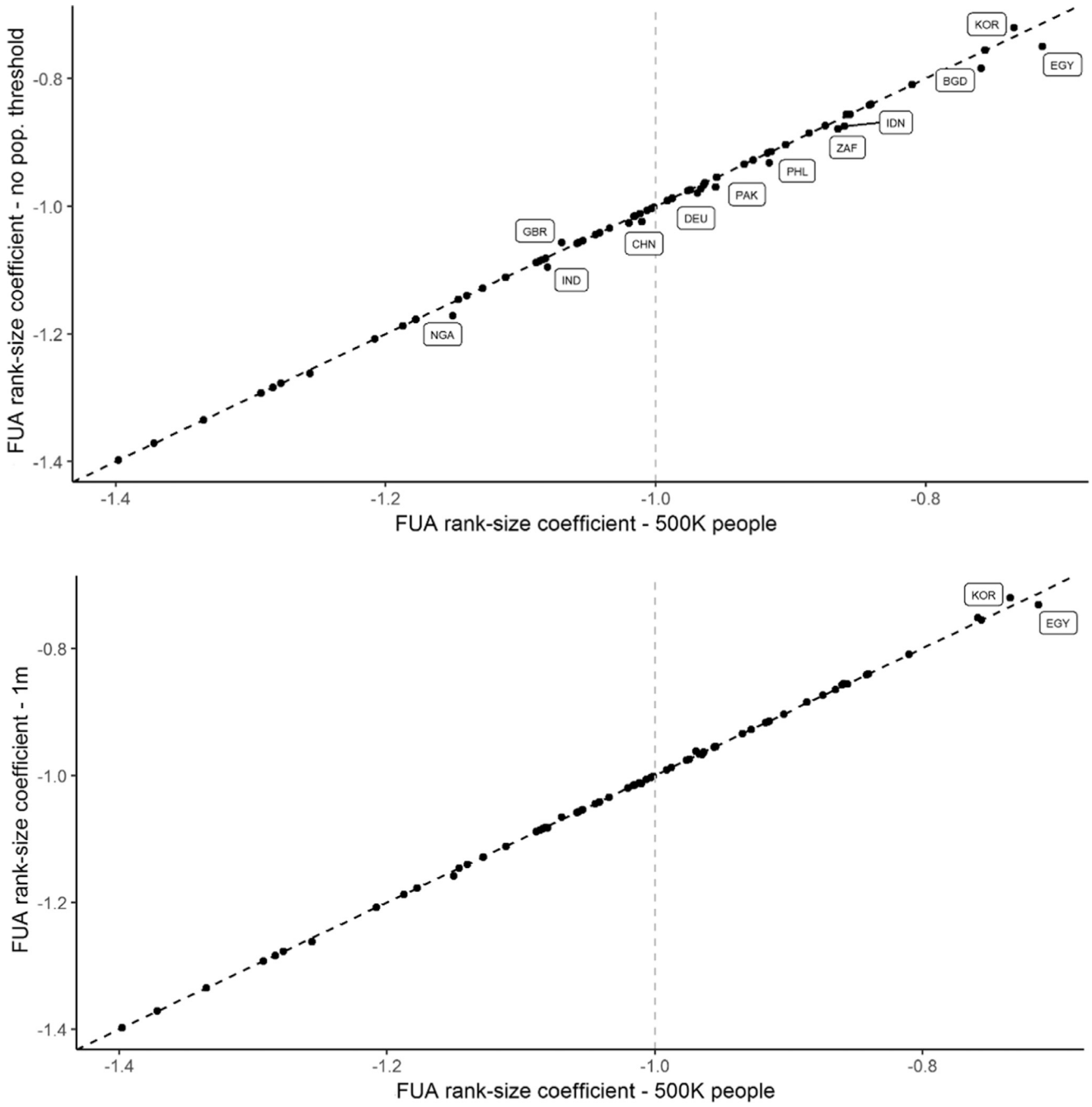

Fig. E.1. Rank-size slope coefficients by urban centre population threshold.
Note: Rank-size coefficient refers to the slope coefficient of a regression of log of FUA rank minus $\frac{1}{2}$ against log of FUA population in 2015 for each country with at least 20 FUAs for all distance thresholds.

Table E.1
Top 5 largest FUAs in Egypt by urban centre population threshold.

Rank	500 K (baseline)	1 million	No threshold
1	23,490,198	32,555,824	36,254,643
2	6,114,874	5,860,478	5,860,478
3	5,860,478	4,753,881	4,753,881
4	2694,394	3698,819	2694,394
5	2468,478	2694,394	2468,478

than 0.01. Imposing a population threshold has a much smaller effect on the rank-size rule slopes, as most cases fall on the 45-degree line.

Nevertheless, the baseline population threshold of 500 thousand people plays an important role in specific cases of high-density and high fragmentation, such as the case of El Cairo (Egypt) illustrated in Table E.1 and Fig. E.2. Without any threshold, El Cairo becomes a FUA of over 36 million people spreading over a large area to the north of Egypt (panel c). The threshold of 1 million partly addresses this but still places the population at 32.5 million people, far above available estimates for Greater Cairo.[21] The baseline case (panel b) comes closer in extent and population to available estimates.

[21] The estimated population of El Cairo (2018) is 20.5 million. Source: https://www.citypopulation.de/php/egypt-greatercairo.php.

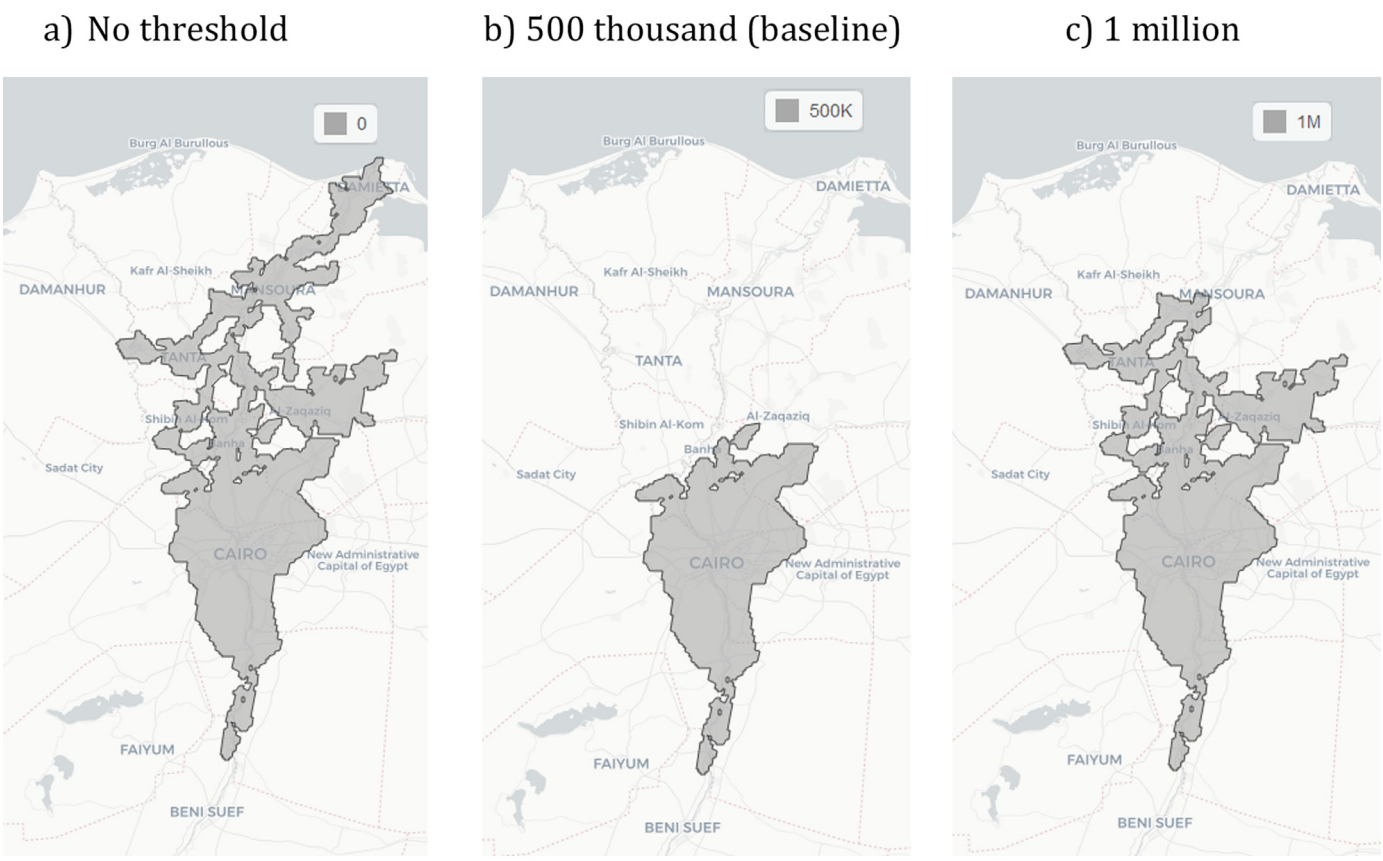

Fig. E.2. El Cairo FUAs by urban centre population threshold.

Annex F. Out-of-sample boundary comparison

Comparing FUAs with alternative functional boundaries

The comparison between alternative (functional) boundaries and FUAs adapted to local administrative boundaries (from here onwards referred to as "administrative FUAs") uses the Jaccard index, a measure of the overlap between any two sets of boundaries. The Jaccard index is the ratio of the population in the area covered by the intersection of two given boundaries (A and B) over the population in the area covered by the union of the two boundaries: $Jaccard\_index = A \cap B / A \cup B$. In the explanation below A refers to alternative boundaries and B to administrative FUAs.

Because it relies on the union of the two boundaries, for a given intersection the Jaccard index can take a low value when A is small compared to B or when B is small compared to A. These cases are qualitatively different if the objective is to assess how well FUAs match alternative functional boundaries. Specifically, a small Jaccard values can arise if a given alternative boundary is relatively small compared to an intersecting administrative FUAs and still only a small percentage of the population in alternative boundaries is not accounted for in administrative FUAs.

To complement the Jaccard index, the percentage of population in alternative boundaries not accounted for in administrative FUAs is defined as $Out\_FUA = A^C/A$, where A^C is the complement of A with respect to the intersection of A and B, i.e., the region of A that is not within B.

Moreover, a measure of how much larger (in % of population in the measured area) administrative FUAs are with respect to a given alternative boundary can be defined as $Excess\_area = 1-(Jaccard\_index + Out\_FUA)*100$, where *excess_area* is equal to zero whenever A (B) is larger than B (A) and B (A) is completely contained in A (B).[22] In all other cases the measure represent the excess of area B (A) with respect to A∪B as a proportion of the area of B (A).

Adapting estimated FUAs into administrative-based boundaries

The criteria to assign a municipality as part of a FUA adapted to local unit boundaries is that 50% of the population of the municipality (measured at the one km^2 level using the GHS population grid) falls within an estimated FUA boundary. Bosker et al. (2019) use the same criteria is used by in their application for Indonesia.[23]

Data preparation for Brazil

First, to adapt FUA to administrative boundaries, boundaries for 5,572 municipal boundaries (2015) covering the entire national territory are used.[24] Applying the aforementioned mapping criteria to 301 estimated FUAs leads to 180 administrative estimated FUAs, out of which 117 include only one municipality.

[22] Demonstration: If there is no area of B outside A: $1-(A\cap B/AUB + A^C/A) = 1-(B/A + A^C/A) = 1-(B + A^C)/A = 1-(A/A) = 0$.

[23] Alternatively, Dingel et al. (2019) construct administrative boundaries based from polygons obtained from nighttime data based on intersection between administrative unit and polygons, with no minimum requirements in terms of land or population coverage. Given the arbitrarieness of administrative boundaries and their differences across countries, it may be more appropriate to considere a minimum population overlap.

[24] The official municipality boundaries are free to download at ftp://geoftp.ibge.gov.br/organizacao_do_territorio/malhas_territoriais/malhas_municipais/municipio_2015/Brasil/BR/.

Table F.1
Summary statistics for *arranjos populacionais* and administrative FUAs in Brazil.

	Number of spatial units	Median population (persons, 2015)	Population range (persons, 2015)	Median area (km^2)	Area range (km^2)
Arranjos	128	250,995	61,467–21,079,374	1,604	187–65,650
Estimated FUAs (administrative-based)	140	208,843	61,467–21,734,127	1,241	117–65,650

Table F.2
Summary statistics for metropolitan areas based on commuting flows (7% threshold) and FUAs in Indonesia.

	Number of spatial units	Median population (persons, 2015)	Population range (persons, 2015)	Median area (km^2)	Area range (km^2)
Metropolitan areas (7% commuting threshold)	37	1,200,156	113,493–31,969,150	2,383	1,020–8,573
Estimated FUAs (administrative-based)	32	1,048,978	64,729–28,432,706	2,254.43	33.7 – 14,952

Next, for as the basis for comparison, the boundaries in the *arranjos populacionais* dataset (2015) are used. This dataset is produced by the Brazilian Institute of Geography and Statistics (IBGE), includes 294 boundaries based on 2015 municipal boundaries, 177 of which have populations of 50 thousand people or more.[25]

Out of 294 FUAs for Brazil, 185 do not cross or overlap any arranjo and are therefore excluded, leading to 140 administrative FUAs as basis for comparison. Meanwhile, 49 arranjos, with a combined population of 3,963,890 persons (2015) are not overlapped by any administrative FUA. These boundaries are not used in the comparison metrics, but their population is added to the total population in arranjos not accounted for in administrative FUAs. Table F.1 summarises the main statistics for the two sets of boundaries.

Data preparation for Indonesia

To map estimated FUAs to administrative areas, we use boundaries for 497 official Indonesian districts (2013). Applying the mapping criteria to 248 estimated FUAs leads to 134 FUAs adapted to local administrative boundaries, out of which 80 include only one district.

For the comparison, metropolitan boundaries defined using the method proposed by Duranton (2015) using a 7% commuting threshold are used.[26] This alternative gives the highest number of separate metropolitan areas (39). In comparison to these metropolitan boundaries, FUAs separate one metro from another even more than the approach based on origin-destination commuting flows, and at the same time do not over-agglomerate into a small number of metros as satellite data-based approaches. In fact, the maximum FUA size for Indonesia is estimated at 28.4 million, below the 32 million using the 7% commuting threshold. See Bosker et al. (2019) for details. Table F.2. summarises the key statistics for the base and comparison boundaries.

References

Ahlfeldt, G.M., Wendland, N., 2016. The spatial decay in commuting probabilities: employment potential vs. commuting gravity. Econ. Lett. 143, 125–129. doi:10.1016/j.econlet.2016.04.004.
Alonso, W., 1964. Location and Land Use. Harvard University Press, Cambridge, MA.
Alonso, W., 1980. Five bell shapes in development. Papers Reg. Sci. Assoc. 45 (1), 5–16.
Bates, D., Mächler, M., Bolker, B., Walker, S., 2015. Fitting linear mixed-effects models using lme4. J. Stat. Softw. 67 (1). doi:10.18637/jss.v067.i01.
Baragwanath, K., Goldblatt, R., Hanson, G., Khandelval, A.K., 2019. Detecting urban markets with satellite imagery: an application to India. J. Urban Econ. doi:10.1016/j.jue.2019.05.004.
Bellefon, M.P., Combes, P.P., Duranton, G., Gobillon, L., Golin, C., 2019. Delineating urban areas using building density. J. Urban Econ. doi:10.1016/j.jue.2019.103226.
Bosker, M., Park, J., Roberts, M., 2019. Definition matters. metropolitan areas and agglomeration economies in a large-developing country. J. Urban Econ. forthcoming.
CIESIN (Center for International Earth Science Information Network), 2017. Gridded Population of the World, Version 4 (GPWv4): Population Count. NASA Socioeconomic Data and Applications Center (SEDAC), Columbia University, Palisades, NY Revision 10 doi:10.7927/H4PG1PPM.
Ch, R., Martin, D.A., Vargas, J.F., 2019. Measuring the size and the growth of cities using nighttime lights. J. Urban Econ. forthcoming.
Chauvin, J.P., Glaeser, E., Ma, Y., Tobio, K., 2017. What is different about urbanization in rich and poor countries? cities in Brazil, China, India and the United States. J. Urban Econ. 98, 17–49. doi:10.1016/j.jue.2016.05.003.
Cheshire, P., 1999. Trends in sizes and structures of urban areas. Hand. Reg.Urban Econ. 3, 1339–1373. doi:10.1016/s1574-0080(99)80004-2.
Corbane, Christina, Florczyk, Aneta, Pesaresi, Martino, Politis, Panagiotis, Syrris, Vasileios, 2018. GHS built-up grid, derived from Landsat, multitemporal (1975-1990-2000-2014), R2018A. Eur. Comm., Joint Res. Centre (JRC) doi:10.2905/jrc-ghsl-10007. PID: http://data.europa.eu/89h/jrc-ghsl-10007.
Corbane, Christina., Pesaresi, Martino, Kemper, Thomas, Politis, Panagiotis, Florczyk, Aneta J., Syrris, Vasileios, Melchiorri, Michele, Sabo, Filip, Soille, Pierre, 2019. Automated global delineation of human settlements from 40 years of landsat satellite data archives. Big Earth Data 3, 140–169. doi:10.1080/20964471.2019.1625528.
Cormen, T.H., Leiserson, C.E., Rivest, R.L., Stein, C., 2001. Section 24.3: Dijkstra's algorithm. In: Introduction to Algorithms. MIT Press and McGraw-Hill, pp. 595–601.
Desmet, K., Henderson, J., 2015. The geography of development within countries. Handbook of Regional and Urban Economics. Elsevier B.V. doi:10.1016/B978-0-444-59531-7.00022-3.
Dijkstra, L., Poelman, H., 2014. A Harmonised Definition of Cities and Rural areas: the New Degree of Urbanisation. European Commission *Regional Policy Working Papers* No. 01.
Dijkstra, L., Poelman, H., Veneri, P., 2019. The EU-OECD Definition of a Functional Urban Area. OECD Publishing, Paris *OECD Regional Development Working Papers*, No. 2019/11 doi:10.1787/d58cb34d-en.
Dingel, J.I., Miscio, A., Davis, D.R., 2019. Cities, lights and skills in developing economies. J. Urban Econ. doi:10.1016/j.jue.2019.05.005.
Duranton, G., 2015. A proposal to delineate metropolitan areas in Colombia. Revista Desarrollo y Sociedad 75, 223–264. doi:10.13043/dys.75.6.
Duranton, G., Puga, D., 2019. Urban Growth and Its Aggregate Implications NBER Working Paper Series 26591 http://www.nber.org/papers/w26591.
Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Handbook of Regional and Urban Economics, 4, pp. 2063–2117. doi:10.1016/S1574-0080(04)80005-1.
Duranton, G., Turner, M., 2012. Urban growth and transportation. Rev. Econ Stud. 79 (4), 1407–1440.
Ellis, P., Roberts, M., 2016. Leveraging urbanization in South Asia. Managing spatial transformation for prosperity and livability. Int. Bank for Reconstruct. Develop. / World Bank Group doi:10.1596/978-1-4648-0662-9.
El-Shakhs, S., 1972. Development, primacy, and systems of cities. J. Dev. Areas 7 (1), 11–36.
Florczyk, A.J., Corbane, C., Ehrlich, D.Freire, S. Kemper, T., Maffenini, L.; Melchiorri, M., Pesaresi, M., Politis, P., Schiavina, M., Sabo, F., Zanchetta, L. (2019): GHSL Data Package 2019, EUR 29788 EN, Publications Office of the European Union, Luxembourg, ISBN 978-92-76-08725-0, doi:10.2760/062975, JRC 117104.
Freire, S., MacManus, K., Pesaresi, M., Doxsey-Whitfield, E., Mills, J., 2016. Development of new open and free multi-temporal global population grids at 250m Resolution. In: *Proceedings of the 19th AGILE conference on geographic information science*, June 14–17. Helsinki.
Freire, S., Schiavina, M., Florczyk, A.J., MacManus, K., Pesaresi, M., Corbane, C., Bokovska, O., Mills, J., Pistolesi, L., Sqires, J., Sliuzas, R., 2018. Enhanced data and methods for improving open and free global population grids: putting 'leaving no one behind' into practice. Int. J.Dig. Earth doi:10.1080/17538947.2018.1548656.
Gabaix, X., Ibragimov, R., 2011. Rank− 1/2: a simple way to improve the OLS estimation of tail exponents. J. Bus. Econ. Stat. 29 (1), 24–39. doi:10.1198/jbes.2009.06157.

[25] The source of the *arranjos populacionais* dataset is: https://www.ibge.gov.br/geociencias/organizacao-do-territorio/divisao-regional/15782-arranjos-populacionais-e-concentracoes-urbanas-do-brasil.html?=&t=o-que-e. Methodological details can be found at: https://biblioteca.ibge.gov.br/index.php/biblioteca-catalogo?view=detalhes&id=299700. For consistency, population values are calculated based on the GHS population grid.

[26] The boundaries were kindly provided by the authors of Bosker et al. (2019).

Galdo, V., Li, Y., Rama, M., 2019. Identifying urban areas combining data from the ground and from outer space: an application to India. J. Urban Econ. doi:10.1016/j.jue.2019.103229.

Géron, A., 2017. Hands-on Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. O'Reilly Media, Inc.

Goh, S., Lee, K., Park, J.S., Choi, M.Y., 2012. Modification of the gravity model and application to the metropolitan Seoul subway system. Phys. Rev. E 86 (2), 026102. doi:10.1103/physreve.86.026102.

Harari, M. (2019). Cities in bad shape: urban geometry in India. *Conditionally accepted, American Economic Review*.

Kim, S., 2007. Changes in the nature of urban spatial structure in the United States, 1890–2000. J. Reg. Sci. 47, 273–287. doi:10.1111/j.1467-9787.2007.00509.x.

Lemelin, A., Polèse, M., 1995. What about the bell-shaped relationship between primacy and development? Int. Reg. Sci. Rev. 18 (3), 313–330.

LeRoy, S.F., Sonstelie, J., 1983. Paradise lost and regained: transportation innovation, income, and residential location. J. Urban Econ. 13 (1), 67–89. doi:10.1016/0094-1190(83)90046-3.

Michaels, G., Rauch, F., Redding, S., 2012. Urbanization and structural transformation. Q. J. Econ. 127 (2), 535–586. doi:10.1093/qje/qjs003.

Muth, R.F., 1969. Cities and Housing. University of Chicago Press, Chicago.

OECD, 2012. Redefining 'urban'. A new Way to Measure Metropolitan Areas. OECD Publishing, Paris doi:10.1787/9789264174108-en.

Pesaresi, M., Ehrlich, D., Ferri, S., Florczyk, A., Freire, S., Halkia, M., Syrris, V., 2016. Operating Procedure For the Production of the Global Human Settlement Layer from Landsat data of the Epochs 1975, 1990, 2000, and 2014. Publ. Off. Eur. Union doi:10.1109/igarss.2016.7730897.

Pesaresi, Martino, Florczyk, Aneta, Schiavina, Marcello, Melchiorri, Michele, Maffenini, Luca, 2019. GHS Settlement grid, Updated and Refined REGIO Model 2014 in Application to GHS-BUILT R2018A and GHS-POP R2019A, Multitemporal (1975-1990-2000-2015), R2019A. European Commission, Joint Research Centre (JRC) PID: doi:10.2905/42E8BE89-54FF-464E-BE7B-BF9E64DA5218.

Schiavina, Marcello, Freire, Sergio, MacManus, Kytt, 2019. GHS Population Grid Multitemporal (1975-1990-2000-2015), R2019A. European Commission, Joint Research Centre (JRC) PID: doi:10.2905/0C6B9751-A71F-4062-830B-43C9F432370F.

Schiavina, M., Moreno-Monroy, A., Maffenini, L., Veneri, P., 2019. GHS-FUA R2019A - GHS functional urban areas, derived from GHS-UCDB R2019A, (2015), R2019A. European Commission, Joint Research Centre (JRC) doi:10.2905/347F0337-F2DA-4592-87B3-E25975EC2C95.

Soo, K.T., 2014. Zipf, gibrat and geography: evidence from China, India and Brazil. Pap. Reg. Sci. 93 (1), 159–181. doi:10.1111/j.1435-5957.2012.00477.x.

Uchida, H., Nelson, A., 2011. Agglomeration index: towards a new measure of urban concentration. Urbanization and Development: Multidisciplinary Perspectives. Oxford University Press doi:10.1093/acprof:oso/9780199590148.003.0003.

UN DESA (2019). *World Urbanization Prospects: The 2018 Revision*, UN, New York, doi:10.18356/b9e995fe-en.

Veneri, P., 2018. Urban spatial structure in OECD cities: is urban population decentralising or clustering? Pap. Reg. Sci. 97 (4), 1355–1374. doi:10.1111/pirs.12300.

Veneri, P., 2016. City size distribution across the OECD. Does the definition of cities matter? Comput. Environ. Urban Syst. 59, 86–94.

Weiss, D.J., Nelson, A., Gibson, H.S., Temperley, W., Peedell, S., Lieber, A., Hancher, M., Poyart, E., Belchior, S., Fullman, N., Mappin, B., 2018. A global map of travel time to cities to assess inequalities in accessibility in 2015. Nature 553 (7688), 333. doi:10.1038/nature25181.

Zhang, Q., Seto, K.C., 2011. Mapping urbanization dynamics at regional and global scales using multi-temporal dmsp/ols nighttime data. Remote Sens. Environ. 115, 2320–2329. doi:10.1016/j.rse.2011.04.032.