TOWARDS A GLOBAL STANDARD TO DEFINE CITIES AND RURAL AREAS

By Lewis Dijkstra[1] (corresponding author, Lewis.Dijkstra@ec.europa.eu), Aneta J. Florczyk, Sergio Freire, Thomas Kemper, Martino Pesaresi and Marcello Schiavina

Abstract

This article presents a new methodology, the degree of urbanisation, and its application to the globe. The degree of urbanisation relies on a population grid to classify local units into three classes: cities, towns & suburbs, and rural areas. These three classes can be further disaggregated into cities, towns, suburbs, villages, dispersed rural areas and mostly uninhabited areas. The population share in rural areas as defined by the degree of urbanisation is similar to the share reported based on national definitions in most countries in the Americas, Europe and Oceania, but radically different in many African and Asian countries. A possible explanation for these differences is that in Africa and Asia smaller settlements are considered rural, while they are classified as urban in the rest of the world. An analysis of the national definitions reported to the United Nations and attempts to replicate these nationally defined shares using density and population size criteria indicate that differences in national definitions (in theory and in practice) make them unsuitable for international comparisons. The global share of population in cities over 300,000 inhabitants, however, is identical between the national definitions and this new method. This implies there is a broader consensus on what is a city than on what is urban and rural. This new definition has been applied to two new global population grids with different methods. In countries where population data is available for very small spatial units, the results are almost identical. In countries where population data is only available for very large units, the results vary substantially. The discrepancies are particularly wide in several African countries. Nevertheless, both grids classify a substantially larger share of the global population as living in urban areas (defined as settlements of 5,000 inhabitants or more) than what is reported by the UN based on heterogeneous national definitions and conventions.

[1] This document reflects the views only of the authors and the European Commission cannot be held responsible for any use which may be made of the information contained therein.

Table of Contents

1. INTRODUCTION

The UN Sustainable Development Goals (SDG) have a stronger subnational focus, including on cities, urban areas and rural areas, than the millennium goals. Whereas, the eight millennium goals could all be measured at the national level, SDG 11 is explicitly targeted at cities and communities. In addition, many of the SDG indicators should be measured not only at the national level, but also for individual cities and for urban and rural areas. This reflects a growing awareness that cities, urban and rural areas present different opportunities for sustainable development and are faced with different problems. Subnational indicators also bring statistics closer to people's daily lives. National averages can obscure the variation within a country. On average, air quality may be quite good in a country, but it may be very poor some of its cities. On average, access to education may be high, but it may be low in some of the rural areas.

Despite the stronger focus on cities, urban and rural areas, the SDGs do not propose a harmonised definition of these types of territories. This creates a risk that even when indicators are measured in an identical manner, they are not comparable because they are applied to territories that are defined in a different way. Several of the SDG 11 indicators are highly sensitive to where the boundary of a city is drawn. For example, access to public transport tends to be higher in the city centre than it is on the outskirts of a city. A city boundary that excludes those outskirts will make the access to public transport seem much higher than if those outskirts were included. The same is true for many of the rural area indicators. For example, the share of population within 2 km of an all-season road will be much higher if settlements with up to 100,000 inhabitants are defined as rural, as is the case in China, than if only settlements with less than 5,000 are defined as rural, as is the case in India.

The emergence of a new statistical tool, the population grid, has created new opportunities to define territories across the globe in a more manner. One benefit of the population grid is that it uses spatial units of the same shape (squares) and size across the entire world, whereas census units have hugely varying shapes and sizes both within and between countries. It also allows us to identify settlements, i.e. population centres, directly without having to rely on other indicators.

This paper argues and demonstrates that the data based national definitions of what constitute urban and rural areas as reported to the United Nations (UN DESA, 2018) are not suitable for international comparisons. Although the UN World Urbanization Prospects clearly indicate that this data is based on national definitions and conventions and list a description of them in the annex, many scholars and journalists have taken this data as sufficiently harmonised to use for cross country comparisons and global assessments. For example, the coming massive wave of urbanisation which has been much discussed (Gross, 2016) so is purely based on data using national definitions.

To encourage scientists to shift to a new paradigm (Kuhn, 1970), however, pointing out the flaws of the current approach is not enough. A new approach that overcomes these flaws is needed. That is why this paper presents a harmonised definition, the degree of urbanisation. In addition, it applies it to a new global, free and open population grid, which reveals a considerably different picture of global urbanisation (see also Melchiorri et al., 2018). Some uncertainty remains, as the quality and spatial resolution of the population data available for some countries is still quite low. Fortunately, more and more statistical offices see the value of producing a population grid based on a geo-coded census or a geo-coded population register. The upcoming census round will allow these estimates of urban and rural population to become more accurate.

The paper is structured as follows. The first section analyses the current national definitions of urban and rural areas, based on definitions reported to the UN and listed in the World Urbanization Prospects (UN DESA, 2018). The second section describes the degree of urbanisation and the data sources used to apply this to the globe. The third section compares the results from the two sources, first for the split between urban and rural areas and secondly for the cities of more than 300,000 inhabitants. To assess the uncertainty of these results, the same method is applied to a second population grid and the difference are assessed.

The last section concludes that to engage in reliable and robust international comparisons a paradigm shift is needed away from using exclusively national definitions and towards a global, people-based harmonised approach. This global definition is not intended replace national definitions, which have the benefit that they can consider much more information than is available globally, but to complement the national definitions so that urbanisation can the situation in cities, urban and rural areas can be compared across national borders in a reliable manner.

The work reported here has been done in the framework of an international voluntary commitment to develop a global, people-based definition of cities and settlements which was launched at the UN-Habitat III conference in 2016. The development of this new definition is a joint project of the European Commission, the Food and Agriculture Organisation (FAO), the OECD, the World Bank and UN-Habitat. The goal of this commitment is to present this definition to the UN Statistical Commission for discussion and, hopefully, approval in March 2020. Within the framework of this voluntary commitment, another linked definition is being tested at the global level, the EU-OECD Functional Urban Area definition, which is also presented in this special issue by Paolo Veneri et al.

2. THE DIFFERENT TYPES OF NATIONAL DEFINITIONS

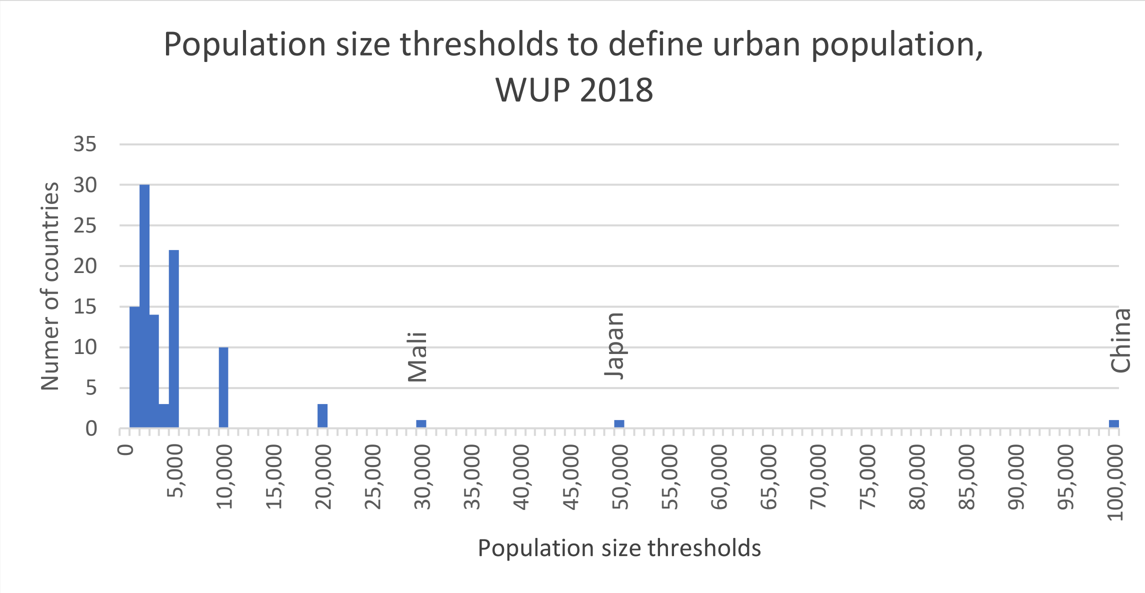
2.1. Definitions using population size

The World Urbanization Prospects (UN DESA, 2018) reports the population share in urban and rural areas in 233 countries and areas. About half of the definitions described in the methodological annex (118) to classify areas as urban, include a minimum population size, either exclusively or in combination with other indicators or criteria. A specific size threshold is mentioned in this annex for 100 countries. Of these, the vast majority (85%) use a population threshold of 5,000 or less (see Figure 1). The most common thresholds are 5,000 (27 countries) and 2,000 (24 countries). Japan and China are outliers with thresholds that are ten to twenty times higher, respectively 50,000 and 100,000. For a good overview of how this has changed over time see (Buettner, 2015).

The impact of a population size threshold depends on the size of the spatial units used. If the units are small in area, many places will drop below this size threshold including small units in a city. If the units are very large in area, many will surpass the threshold including in rural areas. This statistical distortion linked to the shape and scale of the spatial unit is a classic problem known as the modifiable areal unit problem (Gehlke and Biehl, 1934).

Population density is highly sensitive to the size of the spatial unit, which is probably why relatively few countries use it. Only 17 countries are reported as using density as a criterion. Only for 10 countries is the actual density threshold reported. It varies from 150 inhabitants per km[2] in Germany and 200 in Cambodia to 1,500 in China and the Seychelles. In addition, census enumeration areas tend to be small in areas with a high population density and large in area with low population density, which leads to distortions at both end of the density spectrum.

Figure 1 Population size thresholds to define urban population



2.2. Municipalities, localities and settlements

A big obstacle to defining cities and settlements is the lack of consistent data with a high spatial resolution. The UN census recommendations underline that localities should not be equated with the smallest spatial units because a spatial unit can contain multiple small localities and a big locality can be spread across multiple spatial units.

> Localities as defined above should not be confused with the smallest civil divisions of a country. In some cases, the two may coincide. In others, however, even the smallest civil division may contain two or more localities. On the other hand, some large cities or towns may contain two or more civil divisions, which should be considered as segments of a single locality rather than separate localities. (para 2.79)[2]

In other words, settlements (or localities) should be defined independently from civil or administrative divisions. For example, Finland defines an urban area as a population settlement of at least 200 inhabitants, where the distance between residential buildings is no more than 200 meters[3]. In this definition, the first step is to create clusters of residential buildings and only then to count population. It does not directly measure the clustering of population, because historically the data on buildings had a higher spatial resolution than the population. A cadastral map with the outline of each building has a spatial resolution of a few meters, while the resolution of population data varied with the size of spatial unit which range from less than one square kilometer to several thousand square kilometers.

[2] https://unstats.un.org/UNSD/Demographic/sconcerns/densurb/densurbmethods.htm
[3] http://www.stat.fi/ajk/tiedotteet/v2008/tiedote_001_2008-01-15.html

The UN recommendation defines a locality as a distinct population cluster (para 2.78). If the exact location of the population is known, there is no need to make a detour to measure the distance between residential buildings to map population clusters. With growing use of geo-coded censuses, geo-referenced population registers, the accuracy of population data is much higher which allows the direct identification of population clusters.

2.3. Definitions relying on administrative designation

About half (114) the definitions described in the methodological annex use an administrative designation, either exclusively or in combination with other indicators. For international comparisons, the drawback of using administrative designations is that they cannot be applied to other countries. In other words, half of the definitions are not replicable and call into question the comparability of such definitions.

Administrative designations vary. Some list a number of local authorities, as for example Trinidad and Tobago does. Some have an administrative rule. Brazil, for example, requires that every municipality or district, no matter how small or low density, has an administrative centre that is defined as urban. Others combine an administrative designation with a more statistical definition. For example, Zimbabwe's definition includes both places officially designated as urban and places with 2,500 inhabitants or more whose population resides in a compact settlement pattern and where more than 50 per cent of the employed persons are engaged in non-agricultural occupations.

2.4. Other criteria: agricultural employment, infrastructure and services

Three other requirements appear frequently in urban and rural definitions: agricultural employment (37 definitions), certain types of infrastructure (19) and certain services (17). The biggest drawback of such definitions is that they can give rise misleading conclusions. Based on these type of requirements, urbanisation (i.e. an increase in the share of people living in urban areas) can occur without any population growth or movement. For example, in India a settlement can only be urban if less than 25% of the male working population is employed in agriculture. So a settlement of 500,000 can become urban without any population growth or movement, just because its agricultural employment drops below 25%. The same is true of the presence of specific types of infrastructure or services. A settlement can become urban when using such definition merely by the provision of infrastructure or services without any population change.

Given that the urbanisation is typically understood as involving population change, which in turns necessitates further investments in infrastructure and service provision, these definitions can be misleading. Taking the example of infrastructure provision, if in a country becomes more urban because it has provided more infrastructure to multiple settlements, it actually needs less investment in urban infrastructure, but its growing share of population in urban areas will be used to argue the opposite.

In addition to these conceptual differences in defining urban and rural, these three criteria also face two empirical obstacles: big differences between countries and over time, and a lack of harmonised data that can be compared across countries.

Agricultural employment is monitored on annual basis at the national and often the sub-national level and is typically included in a census. Therefore, data availability is not the issue here, but the big differences make it difficult to use a singly threshold in all countries or using the same threshold over time. For example, the global employment share in agriculture dropped from 43%

in 1991 to 26% in 2017 (International Labour Organization, 2018) . Given the magnitude of this reduction, using a single threshold over time would lead to rapid reductions in rural areas (without any population change). From a global perspective, the biggest challenge, however, is the differences between countries. In high income countries, the share of agricultural employment in 2017 was 3% compared to 68% in low income countries. India uses a threshold of 25% agricultural employment, Zimbabwe uses 50%. Applying these thresholds to the high-income countries would mean that they would become entirely urban. Also the other two criteria are likely to classify all of the high-income countries as urban, because access to infrastructure and services tends to be high in almost all of their rural areas, at least from a global perspective.

The availability of harmonised data is a big obstacle to using infrastructure or services in a global definition of urban and rural areas. Defining and measuring infrastructure in a harmonised manner and with high a spatial resolution across the globe will take time. For example, even a relatively simple indicator such as access to an all-season road cannot yet be mapped across the global. The UN sustainable development indicator 9.1.1. (rural population living within 2km of an all-season road) is classified as Tier III, meaning ‘no internationally established methodology or standards are available.’

2.5. Circular reasoning

Including infrastructure in the definition of urban areas also highlights another risk: circular reasoning. If rural areas are defined by the lack of infrastructure, then by definition access in rural areas is going to be low. It actually becomes misleading, to measure access to infrastructure in rural areas. A rural area that would get better infrastructure would cease to be rural. Thus, infrastructure investments in rural areas could lead to a deterioration of the measured access to infrastructure in rural areas because the improved areas would no longer be rural and excluded from this indicator. This risk of a circular reasoning is quite common. If a rural area is defined as a poor area, by definition poverty is going to be higher in rural areas and one can no longer monitor rural poverty.

If a rural area is defined by the share of employment in agriculture, the link between urbanisation and industrialisation becomes tautological. A statement such as ‘Urbanisation leads to lower employment shares in agriculture.’ could be translated as: ‘More areas with a low share of employment in agriculture lead to lower shares of employment in agriculture’. This circular reasoning, however, has not prevented an entire cottage industry arising that analyses and comments the relation between urbanisation and industrialisation.

Definitions that rely mostly or exclusively on the presence of buildings also create a risk of circular reasoning. The amount of built-up land per capita is much higher in rich than in poor countries. For example, cities in North America have about 400 sq meters of built-up land per inhabitant, compared to around 170 in Europe, 75 in Asia and 50 in Africa (data is based on the global human settlement and cities defined by the degree of urbanisation). Therefore, the link between urbanisation and the level of development would be at least in part spurious as built-up area based definitions of urbanisation have a built in pro-rich-country bias.

Using buildings to define population concentrations is also prevalent in research. For example, the Atlas of Urban Expansion (Angel et al., 2011) defines a city purely based on the share of built-up area in cells of 30m by 30m and their proximity to other built-up cells. The global urban footprint of the German Aerospace Centre (DLR) creates clusters of built-up area starting with cells of 12m (Esch et al., 2013).

2.6. Empirical evidence that the national definitions are radically different or that urban areas are *lost in translation*

One could argue that the exact definitions do not really matter that much and that implicitly there is an agreement on what is a city and what is a rural area. To verify if the different national definitions implicitly share a similar character, we measured what density threshold reproduces the same share of urban population as the national definition. This makes the relatively plausible assumption that urban areas are denser than rural areas.

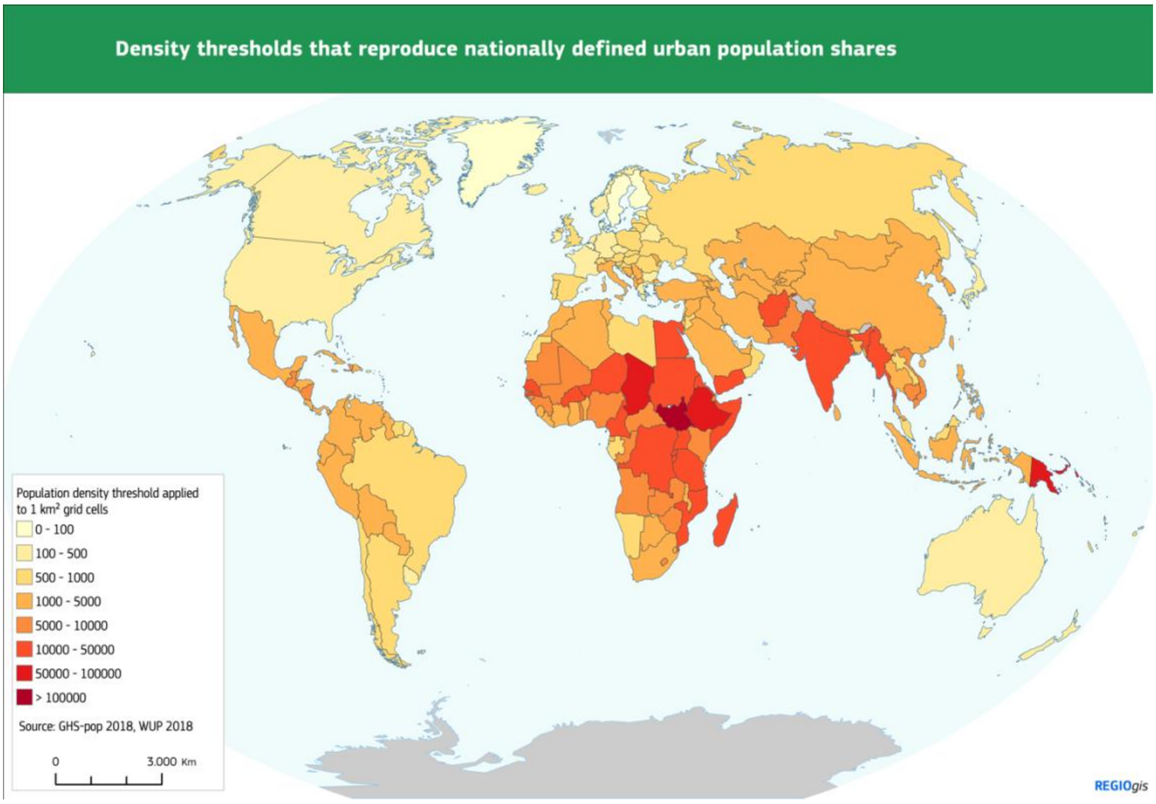
Let us compare the USA with India. They use a similar minimum population size. The USA uses a population size threshold of 2,500 and India uses a population size threshold of 5,000, among other criteria. In the USA 82% of population lives in an urban area, 82% of the population also lives in 1 sq km grid cells with a density of at least 222 residents per sq km. Based on the national definition, India only has 33% of its population in urban areas. To capture only 33% of the Indian population, however, one has to apply a density threshold of 16,705 residents per square km. This is 75 times higher than the USA threshold. Applying this density threshold to the USA would only classify 1.5% of its population as urban. Map 1 shows the density thresholds for each country that reproduces the same population share as the urban share when applied to grid cells of 1 sq km.

The size of the unit has an influence on the density threshold. Larger units have lower average densities. Therefor to capture the same share of population, the density threshold of a larger unit will have to be lower. For example, using cells of 5 by 5 km, the density threshold for the USA drops from 222 to 104 and for India it drops from 16,705 to 4,220. The Indian threshold is still 40 times higher. When using 10 by 10 km cells, the threshold drops to 77 for the USA and to 2,121 residents per sq km in India, still 28 times higher than in the USA. Applying these density thresholds to cells of 5 by 5km and 10 by 10 km cells in the US would lead to classifying only 6% and 13% as urban respectively.

A general picture emerges from this analysis. In the Americas, Europe and Oceania, a low density threshold replicates the urban population share based on national definitions, while in Africa and Asia a much higher density threshold is needed. This pattern holds at all three scales (1 sq km, 25 sq km and 100 sq km) with thresholds at least 10 times higher in Africa and Asia. Using the pre-release version of a different global population grid for 2014 (WorldPop) produced similar results.

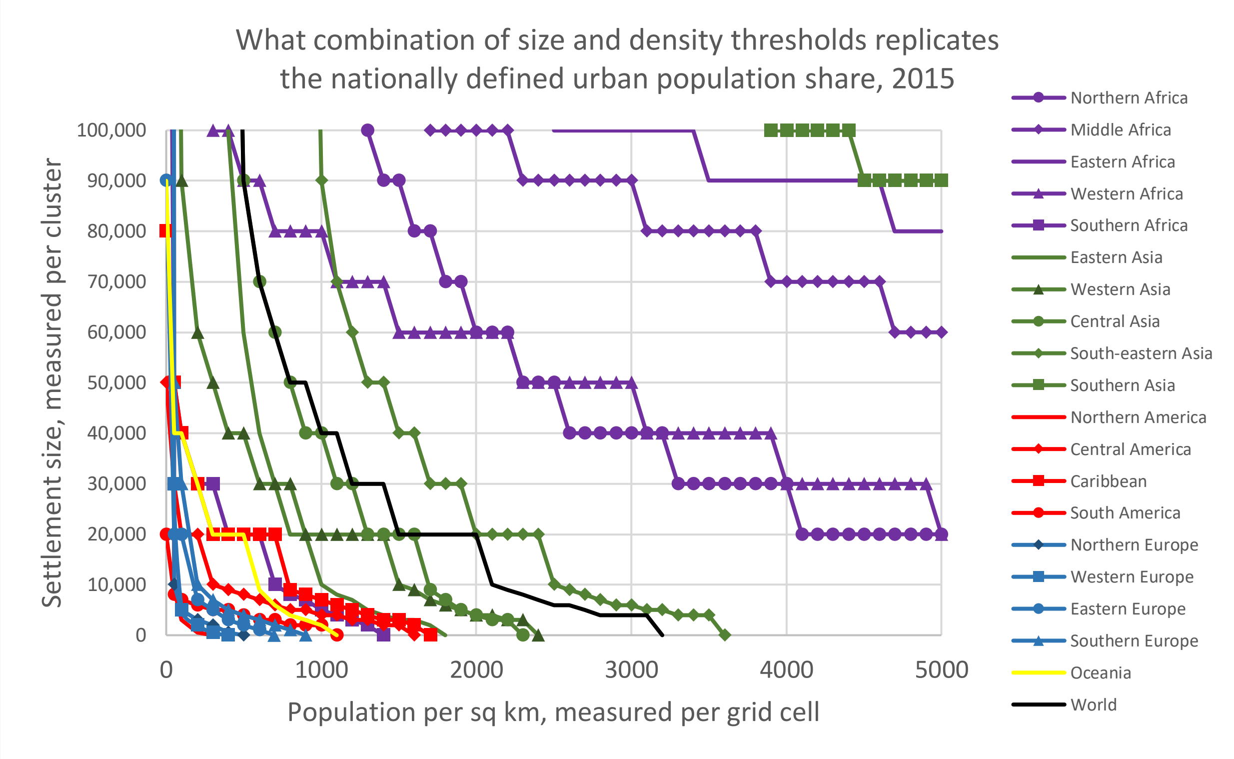
These very large differences lead to two clear conclusions. The national definitions are radically different and are not comparable. Applying a harmonised definition based on population size and density will inevitably lead to very different levels of urbanisation in many countries.

Map 1 Density thresholds that reproduce national defined urban population shares, 2015



We conducted one more test to verify whether using a definition combining population density and population size can reproduce the share of urban population based on national definition. For each sub-region, we identified which combination of population density, applied to a grid cell, and population size, applied to sum of the population in contiguous cell above the density threshold, captures the same share of population.

Figure 2 What combination of size and density thresholds replicates the nationally defined urban population share, 2015



If there was a combination that would work for all the sub-regions, then the lines of all the sub-regions on Figure 2 should cross in a single point. What this test reveals, however, is that in the Americas, Europe and Oceania only a combination of low densities and small population sizes reproduce the national share, while in Africa and Asia only high densities and large population size reproduce the national shares. This supports that hypothesis that urban areas are 'lost in translation', i.e. in some countries, languages or cultures, an urban area refers only to large settlements, while in others it refers to both medium-sized and large settlements.

3. Applying the degree of urbanisation to the globe using a new population grid

3.1. Addressing the modifiable areal unit problem

Ulaanbaatar, the capital of Mongolia, is a city with 1.4 million inhabitants. The area of the municipality is particularly large with 4,700 sq km. As a result, the density of this municipality is very low: less than 300 inhabitants per sq km. Purely relying on municipal densities would inevitably mean that Ulaanbaatar would be classified as rural.

This example captures one of the main reasons why it is so difficult to define a city. This well-known issue is called the modifiable areal unit problem (Gehlke and Biehl, 1934; Openshaw, 1984). The results of a definition or an indicator calculation are dependent on the spatial units used. Using units of a different size or shape can produce very different results. Gerrymandering refers to actively exploiting this to give an unfair advantage to a particular political party when designing voting districts in a first-past-the post system.

Using a population grid reduces the impact of this problem by using units of the same shape and size (squares). The size of these grid cells was selected to be neither too big nor too small. If the squares were too big, smaller settlements would be difficult to detect. Too small, and single settlements may fragment in multiple pieces. For example, the definition of urbanized areas of the US Census bureau is based on census tracts which can be very small. It has to rely on a series of hops, skips and jumps to avoid fragmenting a single settlement (U.S. Census Bureau, 2011).

A last consideration was the resolution at which statistical offices were able or allowed to produce a population grid. Because ideally, this method is applied to official statistics. All European statistical offices were able to produce a grid at 1 sq km resolution, whereas only some were able to produce a finer population grid.

Tests at a 1 sq km resolution showed that in the European context this resolution was neither to big nor too small. It identified both large and small settlements without fragmenting the results.

As explained below the grid concepts are used to classify other spatial units, such as municipalities or enumeration areas. This does reintroduce the modifiable areal unit problem, as the size and shape of these units will determine how closely the population distribution in the grid concepts match the distribution in the three types of administrative of statistical units. In the future more geo-coded data may be produced, but currently the only way to match these areas with data from surveys and administrative sources is by matching the grid cells with administrative or statistical areas that have a name or a postal code.

3.2. The degree of urbanisation as used in the European Statistical System

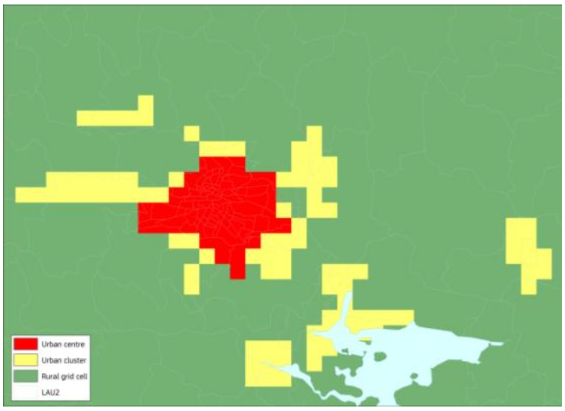
The degree of urbanisation is applied in a two-step process: First the grid cells are defined based on density, contiguity and population size. Subsequently small spatial units are defined based on the type of grid cells the majority of their population resides in.

The degree of urbanisation identifies three types of cells using on a 1 km^2 grid

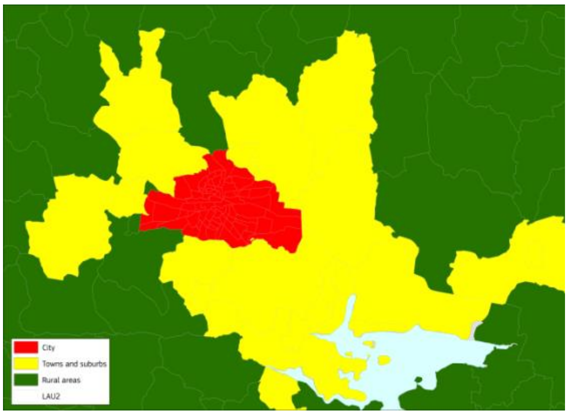
1. **An urban centre** consists of contiguous grid cells with a density[4] of at least 1,500 inhabitants per km^2 and has at least a total population of 50,000; gaps in this centre are filled and the edges are smoothed with an iterative application of the majority rule (if five out of the eight surrounding cells are part of an urban centre, this cell is added to the centre).
2. **An urban cluster** consists of contiguous grid cells with a density of at least 300 inhabitants per km^2 and at least a total population of 5,000; and
3. **Rural grid cells:** grid cells outside urban clusters.

[4] Population density is calculated over the land area of each grid cell. This ensures that cells in a city along rivers, lakes and the sea will be included in the urban centre. This also helps to avoid fragmenting a single city with a wide river running through it.

Map 2: Urban centre, urban cluster and rural grid cells around Cork, Ireland



Map 3: City, towns and suburbs and rural areas around Cork, Ireland



These three types of grid cells are used to classify small spatial units:

1. **Cities** have the majority of their population in urban centres
2. **Towns and suburbs** have the majority of their population in urban clusters, but are not cities
3. **Rural areas** have the majority of their population in rural grid cells.

Urban areas are defined as cities plus towns and suburbs.

This harmonised definition has been implemented by all European national statistical institutes in 2012. Eurostat, the Statistical Office of the European Union, now publishes over 100 indicators by degree of urbanisation. In 2017, the NUTS regulation was amended to include a number of territorial typologies including the degree of urbanisation and the EU-OECD city definition (REGULATION (EU) 2017/2391).

3.3. The global version of the degree of urbanisation

While developing the global version of the degree of urbanisation, two minor changes were made: a simplification and an option.

The simplification: In the European version, the contiguity rule for the urban cluster includes cells located at the corners, while for the urban centre only the four cells directly above, below or next are considered. In short, for urban clusters eight cells are contiguous and for urban centres only four are contiguous. To simplify the method, we opted to use four-point contiguity in both urban centres and urban clusters. The impact of this change is quite limited: it shifted about 1 percentage point of the global population from an urban cluster to rural grid cells.

The option: In a few developed countries, the standard degree of urbanisation method tends to generate multiple urban centres in what functions as a single city. This is typically the case if the country has fairly low-density urban development, a strong separation of land use functions and population data with very high spatial resolution. This occurs, for example, in some cities in Australia and the USA. To remedy this fragmentation, cells that are at least 50% built-up can be included in the urban centre. Such cells typically contain office parks, shopping malls, factories, warehouses and transport infrastructure.

3.4. The refined degree of urbanisation: from three to six classes

The original degree of urbanisation identifies individual cities, but does not provide any distinctions in the other two degrees. The global commitment should define not just cities but other settlements as well. As a result, the degree of urbanisation was refined to also identify smaller individual settlements:

- Towns with a population between 5,000 and 50,000
- Villages with a population between 500 and 5,000

Several languages lack an equivalent term for a town. For example, in Dutch, French, Italian and Spanish, there are terms for cities and villages, but not for a settlement in between the two. As a result, the term had to be translated into 'smaller cities' to make the distinction. This may contribute to some of the confusion surrounding discussion of urban areas and cities.

The degree of urbanisation was modified by using the same density thresholds (1,500 and 300 respectively) but with lower population thresholds:

- A town is defined by an urban centre with a population between 5,000 and 50,000, while a city has an urban centre with a population over 50,000.
- A village is defined as an urban cluster with a population between 500 and 5,000, while a town or suburb has an urban cluster with a population over 5,000.

This approach identifies individual towns and villages, but one type of settlement did not fit either definition but was not a suburb either. These settlements are larger than 5,000 inhabitants (so not a village), have a density below 1,500 inhabitants per sq km (so not a town) and are not adjacent or close to a town or a city (so not a suburb). Therefore, we created two types of towns: dense towns (defined as above) and semi-dense towns, these are defined as an urban cluster of more than 5,000 inhabitants that is located more than 10 km from a city or a town.

To define the towns and the suburbs, three types of grid cells were defined.

- **A dense urban cluster** consists of contiguous cells with a density of at least 1,500 residents per sq km and a population between 5,000 and 50,000 in the cluster
- **A semi-dense urban cluster** is an urban cluster (see above) located more than 10 km from a dense urban cluster or an urban centre (distance is measured between the edges of the clusters)
- **Suburban cells** are the remaining cells in an urban cluster, i.e. not part of a dense or semi-dense urban cluster

Rural areas were broken down into three classes: villages, dispersed rural areas and mostly uninhabited areas.

- **A rural cluster** consists of contiguous cells with a density of at least 300 residents per sq km and a population between 500 and 5,000 in the cluster.

- **Low density rural grid cells** are cells with a density between 50 and 300 inhabitants per sq km

- **Very low density rural grid cells** are the remaining rural grid cells, i.e. those with a density of less than 50 inhabitants per sq km

These grid concepts can be used to define small spatial units as following

- **Towns** have the majority of their population living in a dense or semi-dense urban cluster

- **Suburbs** have the majority of their population living in suburban cells

- **Villages** have the majority of their population living in a rural cluster

- **Dispersed rural areas** have the majority of their population living in low density rural grid cells.

- **Mostly uninhabited areas** have the majority of their population living in very low density rural grid cells.

To ensure that all six classes of the refined degree of urbanisation are properly represented also at the spatial unit level, these units would have to very small. Using large spatial units will lead to lower or even no population in mostly uninhabited areas because few large spatial units would have the majority of their population in very low density grid cells. It also tends to reduce the population share in rural areas.

3.5. The global human settlement layer and its population grid

The Joint Research Centre of the European Commission, in the frame of the Global Human Settlement Layer (GHSL) project, has developed new, global, open and free built-up area and population grids[5] for the epochs 1975-1990-2000-2015, called GHS-BUILT and GHS-POP respectively. The GHS-BUILT grids are the result of fully automated processing of collection of remotely sensed data (Pesaresi et al., 2013). The latest multi-temporal grid (v2017) was produced at 30x30-m spatial resolution by symbolic machine learning method (Pesaresi et al., 2016) from multi-temporal Landsat imagery, using Sentinel-1 derived product (at 20x20-m spatial resolution) as the main training datasets (Corbane et al., 2017).

The population grids combine census-based population data collected by CIESIN at Columbia University (GWPv4.10) with grids reporting built-up densities (Freire et al., 2018). It is currently the most detailed global time series representing residential population distribution[6] and it is free[7]. To produce the population grid, the population within each census unit is disaggregated to the mapped built-up areas (GHS-BUILT) in direct proportion to the built-up density (Freire et al. 2016). This is a very transparent and simple method, but it does have limitations, linked to the detection of built-up areas, the

[5] https://ghsl.jrc.ec.europa.eu/

[6] LandScan by Oak Ridge National Laboratory produces an ambient, not a residential, population grid. The method used to produce this grid is not published and the grid is not free.

[7] It can be downloaded here: https://ghsl.jrc.ec.europa.eu/data.php

distribution of population within the census unit and matching population boundaries to the built-up area layers.

With regard to the detection of built-up areas, there are two risks. Under-detection of built-up areas or omission will lead to an overconcentration of population. Over-detection of built-up areas or commission will lead an overdispersal of the population.

With regard to the distribution of population within a census unit, there are two important issues to take into account. This method will allocate population to non-residential areas as it cannot distinguish between different functions or uses of built-up areas. This will reduce residential densities. This method assumes that the relationship between built-up area and population is identical in the entire census unit. As a result, it will allocate the same number of people to each sq meter of built-up area covered by large villas as to covered by a slum because the built-up layer does not distinguish between different types of residential areas. It will also allocate the same number of people per m^2 of land covered by a 20-story building as covered by a single-story building because the built-up area layer does not capture the height of buildings . The impact of these two issues depends on both the size and the diversity of the census units. If the residential densities are homogeneous within a census unit, this does not lead to a significant distortion. However, if there are many different residential densities and different types of land use within a census unit, they can have a significant impact. In large census units, it will reduce overall densities in residential areas. On average, it will smooth out some of the density differences within a a census unit by increasing the density in low density parts and reducing the density in high density parts.

This method relies on available census data as collected by CIESIN. Notwithstanding significant efforts to harmonize these data, censuses remain heterogeneous in terms quality, accuracy and frequency. Some contain anomalies in the geometry of census units and population counts. In some cases, the boundaries are displaced or population is attributed to the wrong unit, or the counts are inaccurate. Despite the procedures developed to mitigate major anomalies and improve census data (reported in Freire et al., 2018), GHS-POP will inherit most inaccuracies present in this population data and their boundaries.

Despite these issues, this global population grid represents a major step forward in our understanding of the global population distribution and subsequently the level of global urbanisation.

4. Comparing the degree of urbanisation to the figures reported in the world Urbanization Prospects

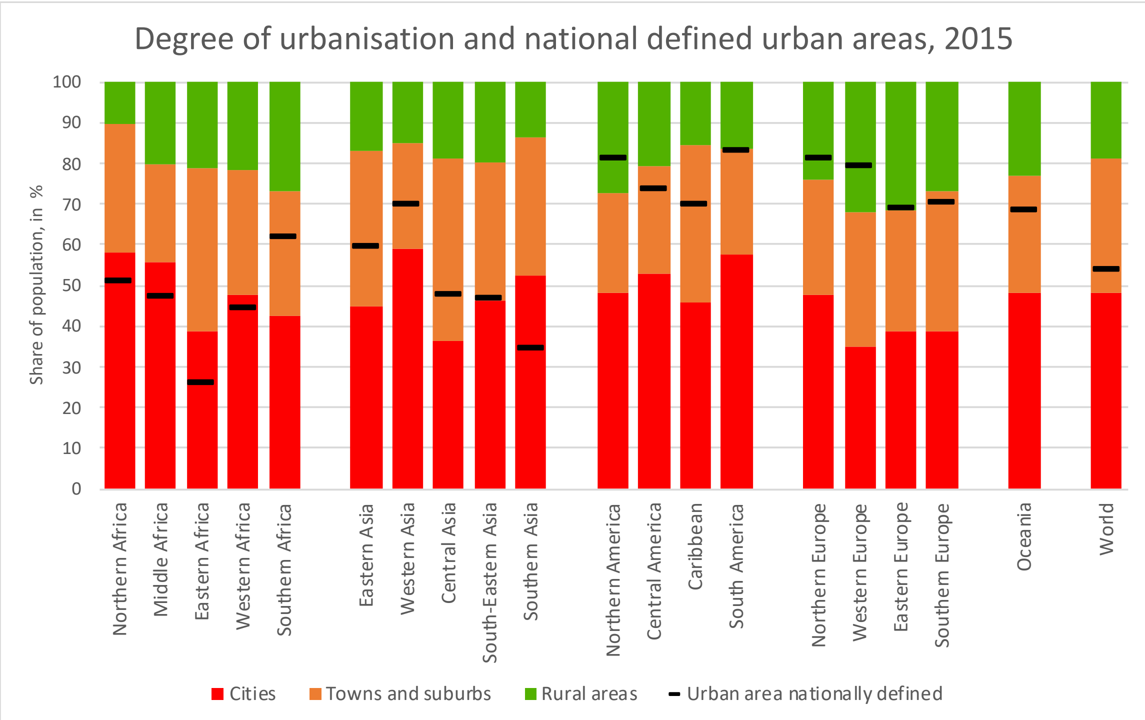
This section compares the results of the degree of urbanisation with the data published by the UN DESA Population Division in the World Urbanization Prospects (UN DESA, 2018) based on national definitions. First, we compare the share of population in urban areas with the degree of urbanisation. Second, we compare cities with at least 300,000 inhabitants with the urban centres (cities) with at least 300,000 inhabitants.

Please note that all the data presented here is only at the grid level. Unfortunately, it was not possible to obtain a global layer with all the census enumeration areas or other small spatial units. Therefore, the data here only covers the first step of the degree of urbanisation: i.e. the coding of the grid cells. For ease of reading, the terms for the spatial units are used here, although the data refers to the grid cell concepts. Finally, we show the results using an alternative global population grid to test the impact of the assumptions needed to create a global population grid.

4.1. Urban and rural population shares

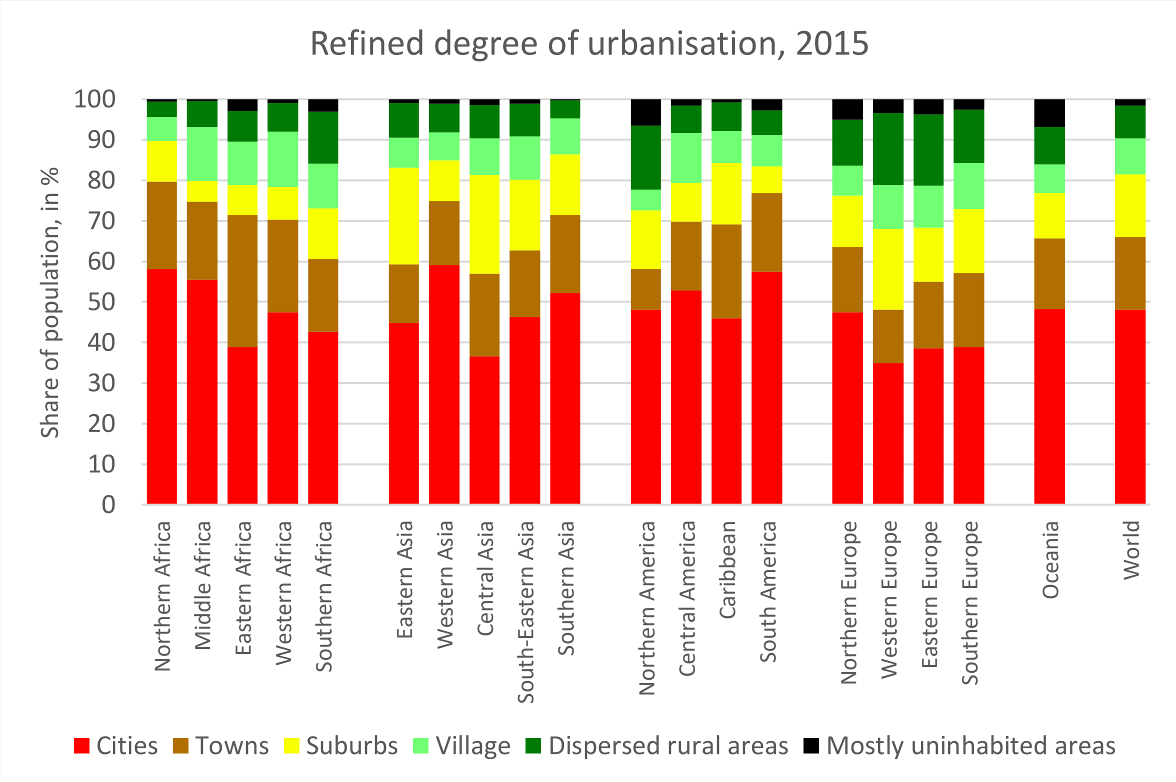
The population shares in the nationally defined rural areas are quite similar to the rural areas as defined by the degree of urbanisation in the Americas, Europe and Oceania (see Figure 3). In Africa and Asia, the population share in nationally defined rural areas is much larger than in the rural areas as defined by the degree of urbanisation. In most cases, it is closer to the population share in cities as defined by the degree of urbanisation. When assessing the national definition, we argued that some countries consider only large settlements (i.e. with at least 50,000 inhabitants) as urban, while smaller settlements are considered rural. For example, Japan and China use a minimum population threshold of 50,000 and 100,000 respectively. China uses a density threshold of 1,500, which is identical to the cities density threshold used in the degree of urbanisation. Part of the difference of the rural population shares, however, is due to the uncertainty in the data (see section 4.3). Given that three quarters of the global population lives in Asia and Africa, the global results also show a much lower share of population in rural areas using national definitions than using the degree of urbanisation.

Figure 3 Degree of urbanisation and nationally defined urban areas, 2015



The three degrees of urbanisation can be further refined to distinguish towns from suburbs and to identify villages, dispersed areas and mostly uninhabited areas in rural areas. The refined degree of urbanisation does require an even more precise population grid. As a result, in sub-regions where the spatial resolution of population data is low the population distribution should be taken as a rough approximation.

Figure 4 The refined degree of urbanisation, 2015



Globally, 18% of the population lives in towns and 15% in suburbs. In Africa, Central and South America suburbs capture a much smaller share of the population than towns. In contrast, suburbs in Northern America, Oceania and in Europe capture a larger or similar share of the population as towns do. In Asia, the picture is less clear with a larger population in suburbs in Eastern and Central Asia and significantly more in towns than suburbs in Western and Southern Asia.

4.2. Comparing cities

In marked contrast to the big differences found in the share of rural population, the share of the population living in cities of at least 300,000 is identical at the global level. The World Urbanization Prospects (2018) lists 1,772 cities with at least 300,000 inhabitants in 2015, which together account for 31% of the world's population. The degree of urbanisation identifies 1,773 urban centres (which defined cities) with at least 300,000 inhabitants, which account for 31% of the global population. This striking near identity is accidental and not something that was reverse engineered. One of the reasons for this high level of agreement could be that conceptually there is less disagreement on how to identify a large city as compared to urban and rural areas.

Despite the identical figures at the global level, substantial variations remain at the national level. With slightly higher shares using national definitions in Europe and the Americas and in some cases lower shares in Africa and Asia.

One conceptual difference is still apparent in these figures. The share of population in cities over 300,000 based on national definitions is considerably higher in North America and Oceania (60% vs 38% and 49% vs 30%). This mainly due to the inclusion of suburbs in the national definitions, while they are reported separately in the degree of urbanisation.

Suburbs have two linked, but distinct characteristics. The first is a hierarchical link to a nearby city or urban area with larger population and a higher density, hence the term sub-urban. The second is a particular style of neighbourhood development that consists primarily of moderate density residential development. While the degree of urbanisation can identify suburban neighbourhoods, it does not identify functional links. The EU-OECD functional urban area definition identifies a commuting zone around each city, which does capture the link between outlying residential areas with a nearby employment centre. A commuting zone, however, does not exclusively contain suburban neighbourhoods but can also encompass smaller towns.

Please note for the nationally defined cities only a point location is provided. As a result, the full extent of these cities is unknown. To compare the two definitions, city points within the boundaries of an urban centre or within a buffer of 5km of an urban centre were considered as matching. Urban centres and city points that were more than 5km apart were considered as not matched. The buffer was added as a precautionary measure to avoid misclassifying matching cities. However, the impact of the buffer is very small as virtually all the points fell within the boundaries within city.

The urban centres included in this comparison have been manually validated. Urban centres that were judged to not represented a city or were not certainly representing a city were excluded from this comparison. As a result, the population in urban centres reported here is slightly lower than in the previous section (45% instead of 48%). Most of the urban centres that were excluded can be found in Middle and Eastern Africa, Southern Asia and Oceania. Population data with a low spatial resolution and buildings which are more difficult to detect using satellite imagery are the main factors leading to urban centres that did not clearly capture a city.

Figure 5 Population in cites over 300,000 based on national definitions and the degree of urbanisation, 2015

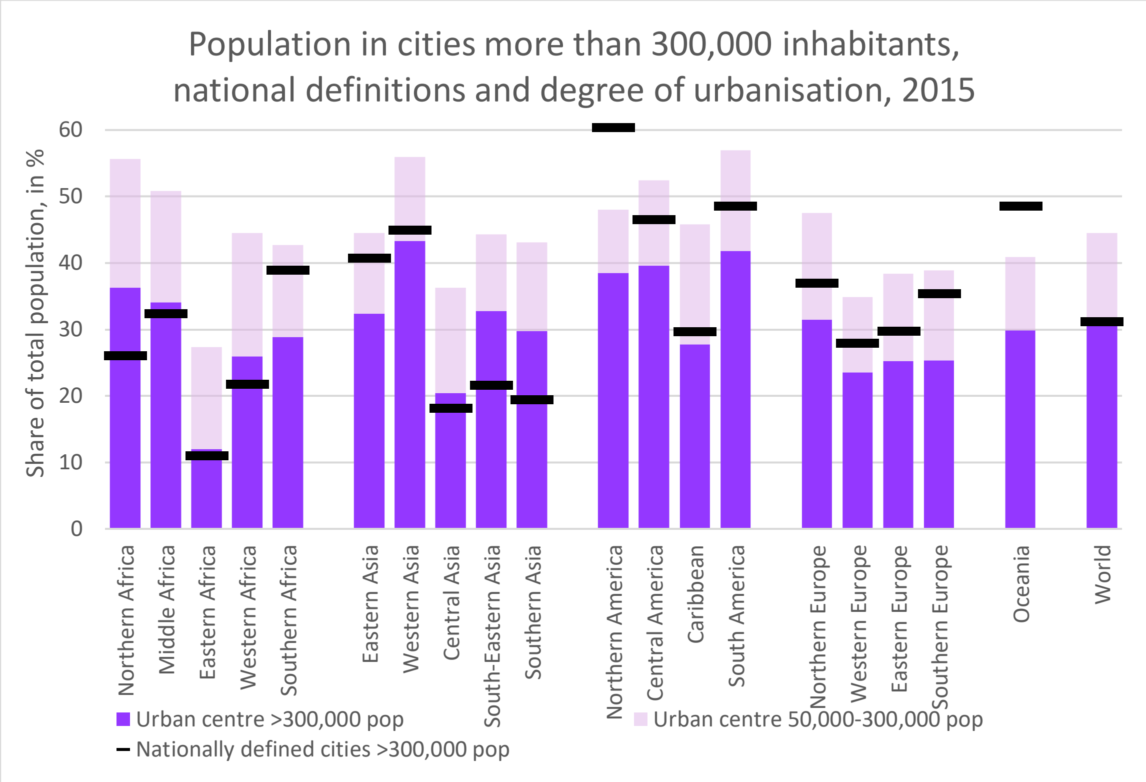
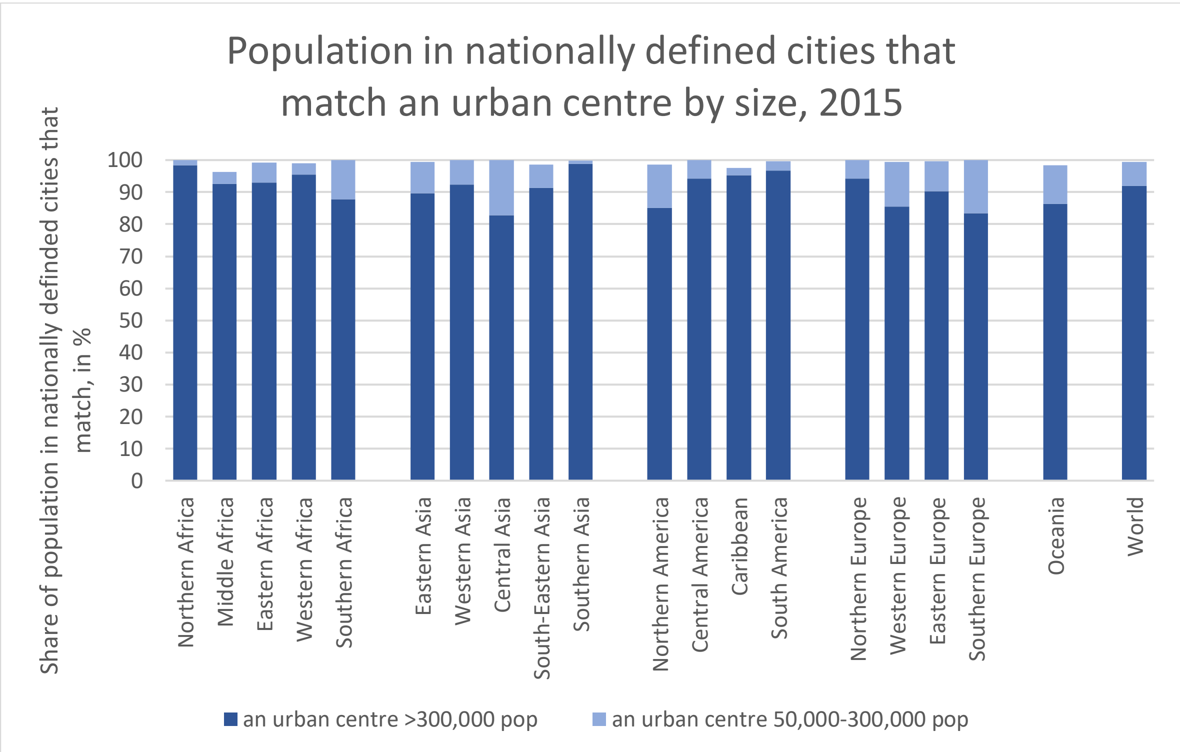
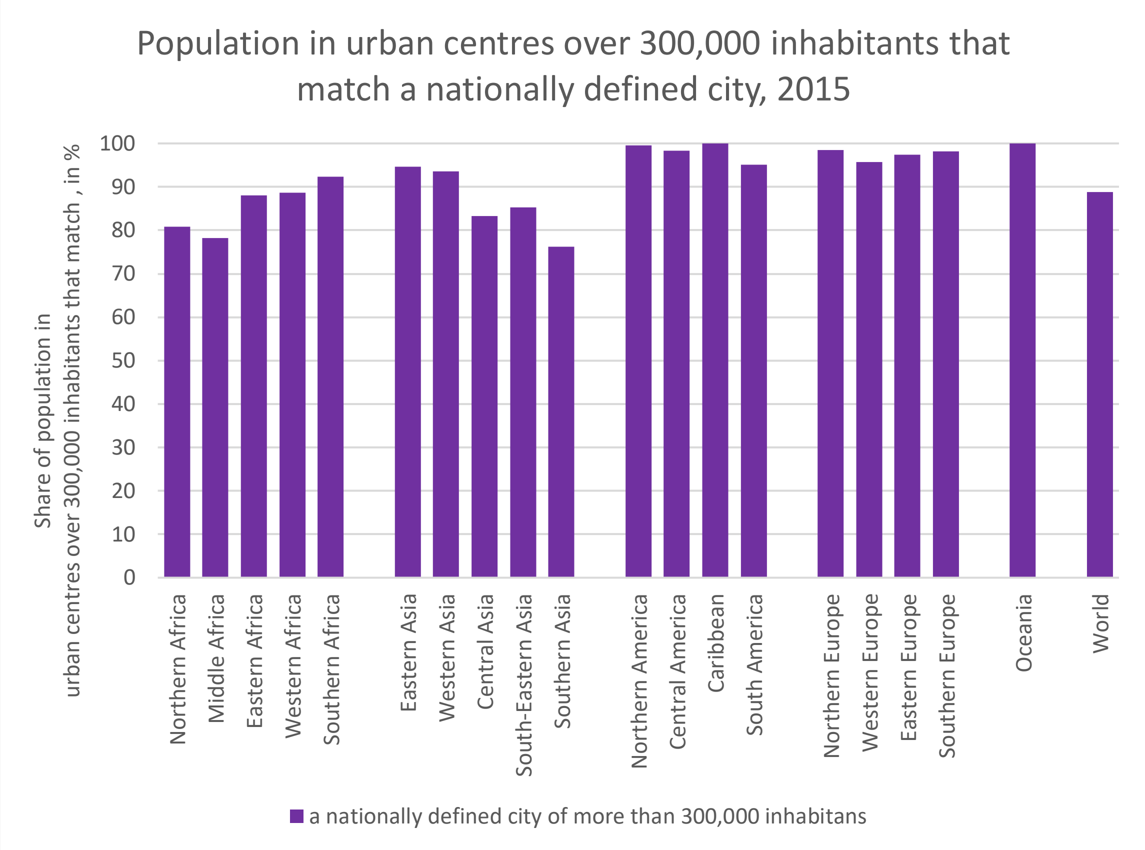


Figure 6 Population in nationally defined cities that match an urban centre by size, 2015



Globally, the degree of urbanisation matches very well with the population in nationally defined cities with at least 300,000 inhabitants. Of the population living in such cities, 92% can be matched with an urban centre of 300,000 inhabitants and another 7.4% city with an urban centre between 50,000 and 300,000. As a result, only 0.6% of the population of these nationally defined cities cannot be matched with an urban centre. The pattern in the various sub-regions is very similar with a strong match with the urban centres over 300,000 and the almost the entire remaining population matching with an urban centre with a population between 50,000 and 300,000.

Figure 7 Population in urban centres over 300,000 inhabitants that match a nationally defined city, 2015



Globally, the urban centres of at least 300,000 inhabitants match well with the nationally defined cities of 300,000 inhabitants with 89% match (see Figure 7). In the Americas, Europe and Oceania the match is particularly high (all over 95%). The match in Africa and Asia is less high, between 76% and 95%. No information is available on smaller nationally defined cities therefore we cannot check whether these unmatched urban centres can be linked to a smaller city.

4.3. Assessing uncertainty

The data presented so far relies on the global, open and free population grid for 2015 produced by combining the CIESIN population data with built-up areas mapped by the Global Human Settlement Layer project. In this section, we compare these results with the results of another draft global population grid (WorldPop) which uses mostly the same population input data but uses a very different approach to disaggregate the population. This grid produced by WorldPop[8] relies on multiple input data to detect built-up areas including the Global Human

[8] http://www.worldpop.org.uk/

Settlement Layer and the Global Urban Footprint as well as other sources such as land cover, land use and transport networks.

One of the key differences between the two grids is that WorldPop will assign population to cells where no built-up areas were detected. Because this leads to a more dispersed population, the share of population in rural areas using the WorldPop grid is consistently higher. One of the benefits of this approach is that it can overcome problems in areas where a significant share of the built-up areas is not detected for example due to tree cover or similar spectral response of the soil and the buildings. By using multiple sources of information, it is less sensitive to errors and limitations of a single source. Some of the drawbacks of this approach are that it is more complicated and can lead to an over-dispersion of the population.

The differences in the degree of urbanisation between the two grids tend to be small for sub-regions where spatial resolution of the data is high (see Table 1). For example, in Northern America the population weighted average size of the census units is 1 sq km. This means that on average it has the same resolution as a 1 sq km grid. These census units do vary in size. In dense areas, they tend to be really small, where they can be individual city blocks. In such an area, creating a 1 sq km grid is only a question of aggregation. In low density areas, census units are bigger, which means some disaggregation is still required to create the grid. Thanks to the high spatial resolution and the high quality and accuracy of geo-spatial information in Northern America, the two grids are very similar and the population shares by degree of urbanisation are virtually identical: the population share in cities using World Pop is 0.4% lower, in towns and suburbs it is 0.3% lower, while in rural areas it is 0.7% higher.

In the sub-regions in Europe, the differences tend to be small with the exception of Eastern Europe, which due to Russia has a much larger size of census units. In the Caribbean, Central and South America, the differences are a bit larger with rural population shares between 9% and 15% higher using WorldPop, due to census units whose population weighted average size by country range from 144 sq km to 939 sq km.

In Asia, the situation is mixed. In Eastern Asia, the differences are not that big and the size of census units is 144 sq km. In the other Asian sub-regions, however, census units are much bigger, between 1000 and 24,000 sq km, and the differences in the rural population share are much bigger, between 9% and 27% higher using WorldPop. In Oceania, the situation is also mixed with very small differences found in Australia and New Zealand, but extreme differences found in Papa New Guinea where the average census unit is 5,000 sq km.

The biggest differences are found in the African sub-region with the exception of Southern Africa. The estimates of the rural population share using WorldPop are between 19% and 40% higher than with using GHS-POP, which is linked to the coarse population data with size ranging from 1,600 to 10,400 sq km.

Table 1 Comparing the population shares by degree or urbanisation using GHS-POP and WorldPop, circa 2015

	Population share according to GHS-POP in %			Population share according to WorldPop in %			Difference in degree of urbanisation using GHS-POP and WorldPop, in % points			CIESIN census units in sq km	
	Cities	Towns & suburbs	Rural areas	Cities	Towns & suburbs	Rural areas	Cities	Towns & suburbs	Rural areas	Population weighted average size	Average size
Northern Africa	58	32	10	48	22	29	-10	-9	19	3,628	5,512
Middle Africa	55	24	20	23	17	61	-33	-7	40	10,400	11,954
Eastern Africa	39	40	21	14	28	59	-25	-12	38	1,684	1,697
Western Africa	47	31	22	25	25	51	-23	-6	29	1,634	2,218
Southern Africa	43	30	27	41	28	31	-1	-3	4	692	752
Eastern Asia	45	38	17	42	36	23	-3	-3	6	145	197
Western Asia	59	26	15	41	26	33	-18	0	18	5,360	3,629
Central Asia	37	45	19	15	39	46	-22	-5	27	23,998	28,271
South-Eastern Asia	46	34	20	35	33	33	-12	-1	13	1,408	1,851
Southern Asia	52	34	14	29	49	22	-23	15	9	956	852
Northern America	48	24	27	48	24	28	0	0	1	1	2
Central America	53	26	21	48	22	30	-5	-4	9	144	97
Caribbean	46	38	16	38	31	31	-8	-7	15	292	277
South America	57	26	17	50	22	28	-7	-4	11	939	1,154
Northern Europe	47	29	24	46	28	26	-1	-1	2	146	173
Western Europe	35	33	32	33	32	34	-1	-1	2	90	31
Eastern Europe	39	30	32	35	23	42	-4	-6	10	5,649	3,481
Southern Europe	39	34	27	38	32	30	-1	-2	3	94	42
Oceania	48	29	23	35	26	39	-13	-3	16	1,095	1,218
World	48	33	19	36	34	30	-13	1	12	1,610	1,653

Correlating the population weighted average size of the census units with the average absolute difference between the degrees of urbanisation shows that more than 50% of the variation is explained by the size (Figure 8). To avoid biasing this analysis, a few outliers were dropped: countries with a population below 50,000 and countries with an average census unit size of more than 10,000 sq km.

A different way to assess the two grids is to see which grid captures most of nationally defined cities, without adding many cities. In about half the sub-regions, the two grids capture the same share of population of national defined cities (Figure 9). In the other half, GHS-POP captures more of the nationally defined cities. For example, in Middle Africa the urban centres based on GHS-POP only miss 7% of the population nationally defined cities, while those based on WorldPop miss 21%. Globally, the difference is small: GHS-POP misses 8%, WorldPop misses 11%.

Conversely, GHS-POP produces globally a higher share of population in urban centres of at least 300,000 inhabitants that do not match a nationally defined city: 11% as compared 5% using WorldPop (see Figure 10). In most sub-regions, however, the results are virtually identical: all the sub-regions in the Americas and Europe, Oceania, Southern Africa and Eastern Asia. In the remaining sub-regions, WorldPop produces less population in urban centres that do not match a nationally defined city. For example, in Middle Africa WorldPop has 5% of the population in urban centres of more than 300,000 inhabitants that do not match a nationally defined city as compared to 11% for GHS-POP.

Figure 8 Differences in degree of urbanisation and the size of census units

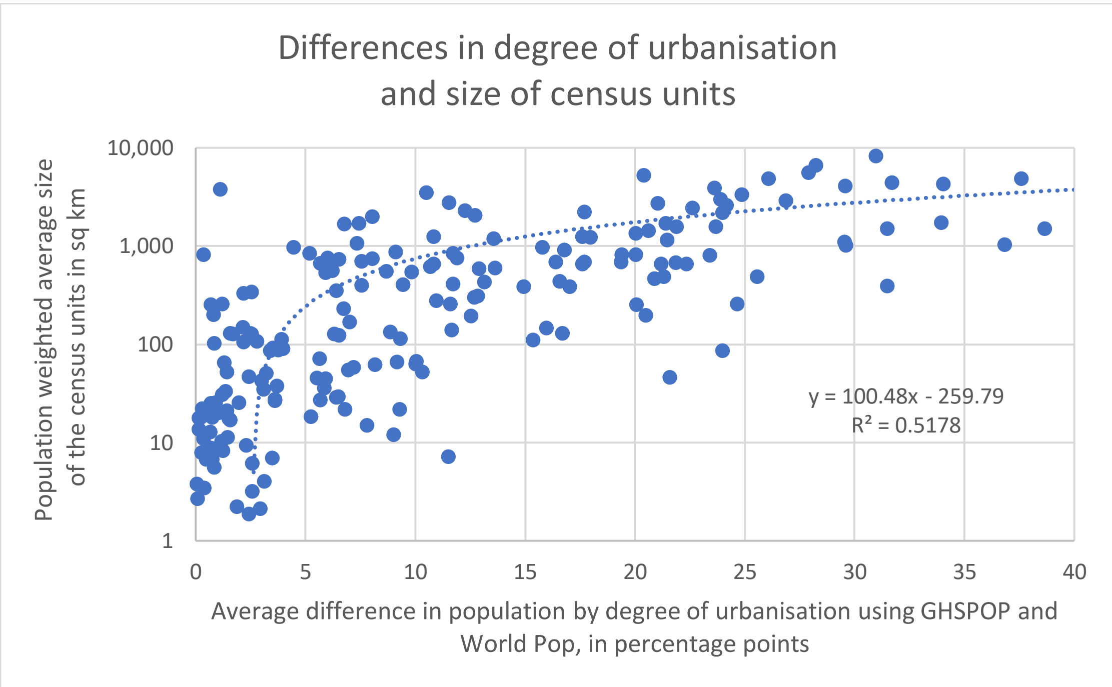


Figure 9 Nationally defined cities that do not match an urban centre of at least 300,000 inhabitants by population grid, 2015

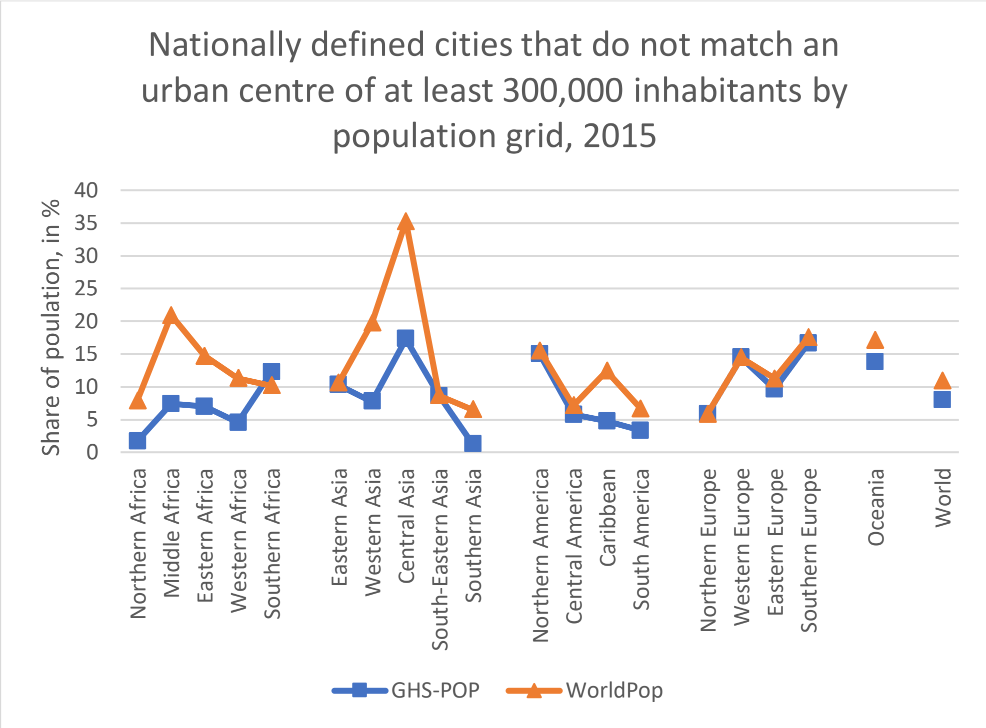
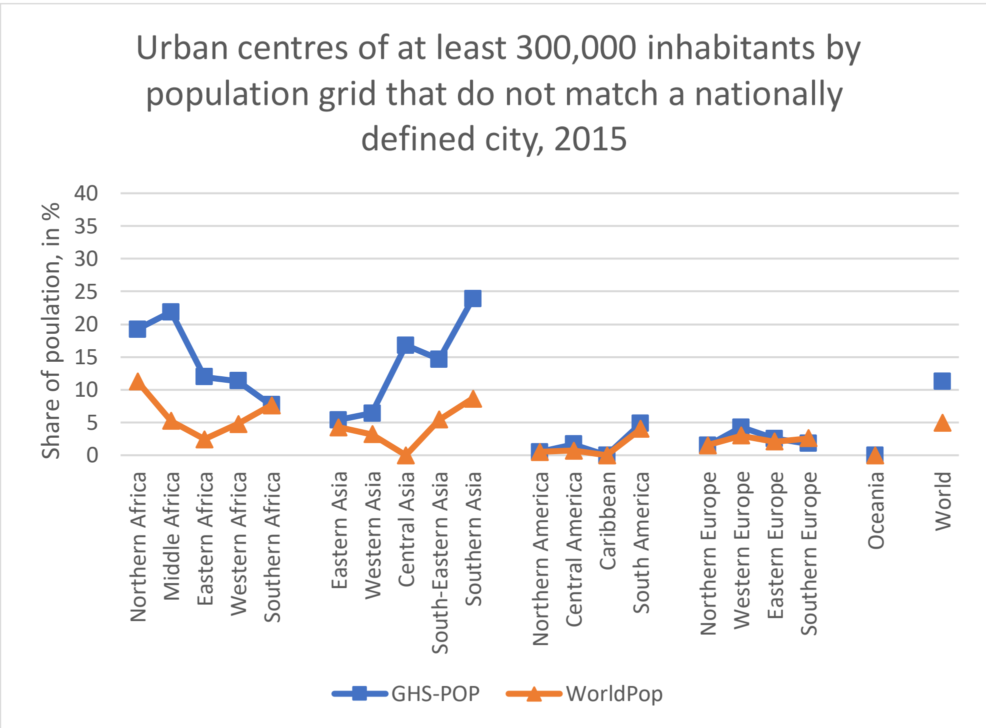


Figure 10 Urban centres of at least 300,000 inhabitants by population grid that do not match a nationally defined city, 2015



The comparison of the two grids does not allow us to identify which one is the closest to real distribution of population. It does show in which locations there is a considerable uncertainty. Notably, in Africa, the two grids produce very different estimates of the rural population (20%

with GHS-POP vs 50% with WorldPop). Better remote sensing data and other geospatial data can help to improve the accuracy of the population grids, but higher resolution population data will be needed to overcome some of the inherent challenges of downscaling residential population. The next round of censuses will produce higher resolution population data and more population grids. In countries where such high resolution data is missing the location of mobile phones could be used to estimate the distribution of the residential population.

Despite the differences between the two results, one conclusion is uncontested the population share in urban areas (defined as settlements over 5,000 inhabitants) is considerably higher than the figures reported in the World Urbanization Prospects 2018 (56% vs 70% with WorldPop and 81% with GHS-POP).

5. Conclusions

This paper argues that national definitions of urban areas are too different to be used for international comparisons. With minimum settlements sizes ranging from 200 inhabitants (Denmark) to 100,000 (China), this should not come as a great surprise. The main theoretical reasons why they are too different are that 1) half the countries rely on an administrative designation, which cannot be replicated, 2) the other half do use a statistical definition but many rely on indicators that are either not available for all countries or not suitable for a global definition and 3) the countries that use a minimum population size use quite different thresholds and apply them to units of very different shapes and sizes. A new global population grid was used to test whether empirically the population shares in nationally defined urban areas could be replicated using criteria of density and size. Multiple test showed that the sub-regions in the Americas, Europe and Oceania required low population density and size thresholds, while those in Africa and Asia typically required high population density and size thresholds. This suggests that the term ‘urban areas’ may be lost in translation. Some countries use it to refer exclusively large settlements, while others use it both from medium-sized and large settlements.

A new method to define 1) cities, 2) towns and suburbs, and 3) rural areas called the degree of urbanisation has been tested at the global level by applying it to a new population grid (GHS-POP). This produces very similar population shares in rural areas as compared to nationally defined rural shares in the Americas, Europe and the Oceania, but very different shares in Africa and Asia. This could be because urban areas in Africa and Asia typically only refer to larger settlements, while medium-sized settlements are included in the rest of the world. The global population share in cities of 300,000 inhabitants or more as defined by the degree of urbanisation is virtually identical to the share based on nationally defined cities of at least 300,000 inhabitants. This suggests that there is a greater consensus on what constitutes a city than what constitutes an urban area. It also implies that the degree of urbanisation does a good job identifying large cities. The degree of urbanisation, which was originally developed for the European Statistical System, was further refined to capture the full settlement hierarchy. The refined version identifies six classes: 1) cities, 2) towns, 3) suburbs, 4) villages, 5) dispersed rural areas and 6) mostly uninhabited areas.

The method and the data in this paper represent a major breakthrough as it is the first time that a comprehensive and harmonised list of cities was created endogenously. Other research so far relies on exogenous lists of cities and settlements, which means their results reproduce all the biases present in the national lists of cities and settlements (for a discussion of these issues see Forstall et al., 2009). This is the case for the World Bank agglomeration index (Uchida and Nelson, 2009), Global Rural-Urban Mapping Project (GRUMPv1) (Balk et al. 2006) and the "Making room for a planet of cities" (Angel et al., 2011).

The degree of urbanisation was also applied to a draft global population grid developed by WorldPop to assess the uncertainty in these disaggregation grids and their consequences for the estimates of the urban and rural population shares. It shows that in particular in Africa and some parts of Asia, a high degree of uncertainty remains and recent population data with a much higher spatial resolution are needed to produce more conclusive urban and rural population shares.

Developing and agreeing on a global, people-based definition of cities and settlements is essential to reliably compare urbanisation levels, rural areas and cities across national boundaries. The UN Sustainable Development Goals call for indicators to be produced for urban and rural areas as well as for individual cities. The degree of urbanisation has been developed and tested to allow all countries in the world to complement their own national definitions with an international one that enables far more reliable global comparisons.

6. References

Angel, S., Parent, J., Civco, D.L., Blei, A.M., 2011. Making room for a planet of cities, Lincoln Institute of Land Policy. https://doi.org/10.4337/9781849808057.00023

Balk, D. L., U. Deichmann, G. Yetman, F. Pozzi, S. I. Hay, and A. Nelson. 2006. Determining Global Population Distribution: Methods, Applications and Data. Advances in Parasitology 62: 119–156

Buettner, T., 2015. Urban Estimates and Projections at the United Nations: The Strengths, Weaknesses, and Underpinnings of the World Urbanization Prospects. Spat. Demogr. 3, 91–108. https://doi.org/10.1007/s40980-015-0004-2

CIESIN (Center for International Earth Science Information Network), 2011. Global Rural-Urban Mapping Project, Version 1 (GRUMPv1): Urban Extents Grid. Columbia University. [WWW Document]. Glob. Rural. Mapp. Proj. Version 1. https://doi.org/10.7927/H4GH9FVG

Corbane, C., Pesaresi, M., Politis, P., Syrris, V., Florczyk, A.J., Soille, P., Maffenini, L., Burger, A., Vasilev, V., Rodriguez, D., Sabo, F., Dijkstra, L., Kemper, T., 2017. Big earth data analytics on Sentinel-1 and Landsat imagery in support to global human settlements mapping. Big Earth Data 1, 118–144. https://doi.org/10.1080/20964471.2017.1397899

Esch, T., Marconcini, M., Felbier, A., Roth, A., Heldens, W., Huber, M., Schwinger, M., Taubenbock, H., Muller, A., Dech, S., 2013. Urban Footprint Processor—Fully Automated Processing Chain Generating Settlement Masks From Global Data of the TanDEM-X Mission. IEEE Geosci. Remote Sens. Lett. 10, 1617–1621. https://doi.org/10.1109/LGRS.2013.2272953

R.L. Forstall, R.P. Greene, J.B. Pick, 2009. Which are the largest? Why lists of major urban areas vary so greatly. Tijdschrift voor economische en sociale geografie, 100 (3): 277-297.

Freire, S., K. MacManus, M. Pesaresi, E. Doxsey-Whitfield, and J. Mills, 2016. Development of New Open and Free Multi-temporal Global Population Grids at 250 m Resolution. 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., R., S., 2018. Enhanced data and methods for improving open and free global population grids: putting 'leaving no one behind' into practice. Int. J. Digit. Earth in press.

Gehlke, C.E., Biehl, K., 1934. Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material. J. Am. Stat. Assoc. 29, 169. https://doi.org/10.2307/2277827

Gross, M., 2016. The urbanisation of our species. Curr. Biol. 26, R1205–R1208. https://doi.org/10.1016/J.CUB.2016.11.039

International Labour Organization, 2018. Statistics and Databases [WWW Document]. Website. URL https://www.ilo.org/global/statistics-and-databases/lang--en/index.htm (accessed 9.5.18).

Kuhn, T.S., 1970. The Structure of Scientific Revolutions, University of Chicago Press. https://doi.org/10.1119/1.1969660

Melchiorri, M.; Florczyk, A.J.; Freire, S.; Schiavina, M.; Pesaresi, M.; Kemper, T., 2018. Unveiling 25 Years of Planetary Urbanization with Remote Sensing: Perspectives from the Global Human Settlement Layer. Remote Sens. 10(5), 768.

Openshaw, S., 1984. Ecological Fallacies and the Analysis of Areal Census Data. Environ. Plan. A. https://doi.org/10.1068/a160017

Pesaresi, M., Huadong, G., Blaes, X., Ehrlich, D., Ferri, S., Gueguen, L., Halkia, M., Kauffmann, M., Kemper, T., Lu, L., Marin-Herrera, M.A., Ouzounis, G.K., Scavazzon, M., Soille, P., Syrris, V., Zanchetta, L., 2013. A global human settlement layer from optical HR/VHR RS data: concept and first results IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing, 6(5):2102-2131. https://doi.org/10.1109/jstars.2013.2271445

Pesaresi, M., Syrris, V., & Julea, A. (2016a). A new method for earth observation data analytics based on symbolic machine learning. Remote Sensing, 8(5), 399. doi:10.3390/rs8050399

U.S. Census Bureau, 2011. Urban area criteria for the 2010 census. Fed. Regist.

Uchida, H., Nelson, A., 2009. Agglomeration Index : Towards a New Measure of Urban Concentration.

UN DESA, P.D., 2018. World Urbanization Prospects 2018 [WWW Document]. URL https://esa.un.org/unpd/wup/ (accessed 9.4.18).