Full Title and Running Title:

A Free, Open Source Tool for Identifying Urban Agglomerations using Polygon Data

Jennifer Day
Faculty of Architecture, Building, and Planning
The University of Melbourne
Room 520, 757 Swanston St.
Parkville, Vic 3010
Australia
+61 451 054 878
jday@unimelb.edu.au

Corresponding author:

Yiqun Chen
Faculty of Engineering
The University of Melbourne
Parkville, Vic 3010
Australia
+61 433 811 913
yiqun.c@unimelb.edu.au

Peter Ellis
Lead Urban Economist
South Asia Urban Development Unit
The World Bank
1818 H Street, NW Washington, DC 20433
USA
pellis@worldbank.org
+1 202 458 2814

Mark Roberts
The World Bank
South Asia Urban Development Unit
1818 H Street, NW Washington, DC 20433
USA
mroberts1@worldbank.org
+1 202 473 5721

Keywords

Urbanisation
Urban extents
Urban agglomeration
Polygon data
Metropolitan region

JEL codes

C81
J11
R58

Total words (not including references, tables, or abstract)

6285

Abstract

This paper describes the function of a software tool for identifying urban agglomerations in low-information settings using free and open data. The framework outlined here is designed to work using polygon data. This paper describes the advantages and disadvantages of using polygon-based geographies in regional analysis, discusses the practical and ethical challenges of distinguishing urban from rural regions, and discusses the relevance of this tool in the analysis of global city regions. It also describes the logical structure of our polygon-based

software tool and directs interested readers to the source code. We finally examine the agglomeration results for Sri Lanka and compare them with published urbanization figures. We conclude that there are very large disparities between our model's outputs and the urbanization estimates from the United Nations, and that our tools can be used as a less-discretionary way to identify actual levels of urbanization. We hope that other analysts will continue to refine the progression toward a less discretionary model of identifying urban regions.

1 Introduction

Analysis of urbanization and urban development are often complicated by non-standard processes for identifying urban agglomerations in different countries and settings. This is particularly problematic in cross-country analyzes, where analysts need urban distinctions that are independent of country political processes.

Official conversion of land from rural to urban often significantly lags behind changes in official designation for a number of practical, bureaucratic and political reasons (J. Day & Ellis, 2013b; J. Day & Lewis, 2013; Rotgé, 2001). The software tool and process we present here responds to a need for a standardized way to identify metropolitan agglomerations in settings where information is limited to basic Census data – a call which is echoed by many other authors and analysts, e.g., (Gamba & Herold, 2009, p. xi; Pesaresi *et al.*, 2013).

This paper introduces a software tool for identifying urban agglomerations in data-poor environments using free, open data. The tool is built in Cran R using data hosted in the public domain and is designed to construct agglomerations using data in polygon format – that is, data attributed to polygon geographies. The resulting agglomerations will generally match administrative polygons, in which Census data and national accounts data are typically reported. This tool and process, which we call the Agglomeration Polygon Algorithm, then, can provide the spatial foundation for analysis of metropolitan areas using Census and national

accounts data. After a review of the relevant literature, this paper discusses the polygon algorithm's inputs, decision logic, and functionality. We then demonstrate the tool's functionality and outputs using data for Sri Lanka for the most recent Census in 2011. Finally, we compare the agglomeration model outputs with published United Nations urbanization figures for Sri Lanka.

2 Background

For analysts attempting to define and delineate urban areas from adjacent rural and contiguous metropolitan areas, there are a number of available approaches. Perhaps the most common is to simply accept official urban designations provided by governments. This can result in significant under-reporting of urban populations where political designations have not kept pace with urbanization (J. Day, Chen, Ellis, & Roberts, 2016), or over-reporting of urban populations. One Census Bureau-defined procedure to classify rural populations in the United States significantly overcounts the urban population, including 30 million rural residents of metropolitan counties (Isserman, 2005). More generally, as Scott (Scott, 1998) points out in his book *Seeing Like a State*, the intentionality behind population accounting is often strategic or political, rather than to provide analysts with useful delineations.

A second general approach to delineate urban areas is to use remote sensing. Various remote sensing options also exist to identify urban extents, and these can be very useful to analysts. Satellite images for nighttime lights are available for most of the Earth (Doll, 2008). However, reliance on these can leave out large swaths of urbanized areas that are not electrified, such as the peri-urban areas of Port Vila, Vanuatu (author's personal experience), or large areas of Pyongyang (Shim, 2014). Urban heat effects and thermal signatures are also possible to map (Lo, Quattrochi, & Luvall, 1997), but may suffer from the same problems, and are also not readily available for the entire world. Also, remote sensing geographies are generally not directly aligned with political geographies, which can be useful in some analysis. However, for analysis requiring the urban boundaries to be aligned with economic or other data, e.g., (J. Day & Ellis, 2013a, 2013b; J. Day & Lewis, 2013), urban extents defined in terms of political geographies, can be useful.

Other remote sensing options distinguish different types of land cover at varying levels of sophistication. Taubenböck *et al.* (2013) use building heights and other building features generated from three-dimensional models of urban regions to identify and delineate central business districts (CBDs) from their surrounds. Their process would be difficult to apply to areas outside of CBDs in major cities

because it requires sophisticated 3D models, which are available in their test cities (London, Paris, and Istanbul) but not for low-information cities like Colombo.

Some analysts also identify processes that simply require spectral analysis of satellite images, which are available for most of the world, e.g., (Bagan & Yamagata, 2014; Møller-Jensen, 2013; Tatem, Noor, & Hay, 2005; Wang *et al.*, 2012; Zha, Gao, & Ni, 2003). Pesaresi *et al.* (2013) have developed a trial of a sophisticated process that defines urban regions from spectral analysis of buildings with population-based datasets used as ancillary validation. This important initiative or others like it will hopefully result one day in a universal set of urbanization maps. Even so, one problem with these types of approaches is that they do not offer a way to distinguish adjacent metropolitan regions from each other or to match identified metropolitan regions with economic and other data; another is that they have not yet produced maps for the entire world, let alone maps for multiple time periods.

A third general approach to designate urban regions is to rely on published spatial data such as population, land use, land cover, and other features provided in available datasets. Uchida and Nelson (2008) define their Agglomeration Index framework and present a method to identify metropolitan agglomerations in data-poor settings. They then demonstrate their Agglomeration Index outputs in a cross section of 210 countries and report national urbanization levels based on

their tool. Uchida and Nelson's algorithm is a useful start, but they do not provide guidance on a number of issues that we found were pertinent to detailed metropolitan analysis. Their process provides no flexibility for selecting seeds that are at an appropriate density for a given country. None of the studies described above provide working software or source code. Also, none provides a framework to distinguish contiguous metropolitan areas, or on how to allocate border regions in two adjacent metropolitan areas, or on how to decide whether a large metropolitan area should be broken into two or more adjacent metropolitan areas.

Contiguous urbanization is a feature present in urban and urbanizing regions from small islands in the Caribbean and the Pacific, to the Nile riverfront, to China's Pearl River Delta – but decision tools to understand contiguous urbanization are rare. Taubenböck *et al*'s (2013) process offers a method to distinguish CBDs from the surrounding urban fabric, and also to distinguish multiple CBDs in a single metropolitan mega-region, but not to delineate contiguous metropolitan areas and assign the urban edges to one region or the other. Isserman (2005) offers a new taxonomy, a *rural-urban density code* designed to correct this problem, which he offers to provide upon request, but provides no editable software code to perform these tasks. Morrill *et al.* (Morrill, Cromartie, & Hart, 1999) find that using smaller geographies than the county (block groups) results in a disaggregation of some larger metropolitan areas into

smaller areas. However, they do not provide a process for how this disaggregation could be performed by other analysts.

Building urban agglomerations from small geographies also means that the resulting agglomerations are not compatible with country Census geographies, as noted in Day *et al* (2016). These smaller-area and point data are useful on one hand, as an analyst is not confined to including entire Census geographies in the analysis when in reality only a portion of the geography is urbanized. However, this data format can be a problem in cases where the analyst would like to incorporate other datasets that are based on Census geographies, such as poverty rates, gross regional domestic produce (GRDP), and many other data that are widely reported in Census geographies.

A variety of spatial clustering processes exist that have not been explicitly designed to identify urban agglomerations, but which could be used for such purposes (Verma, Srivastava, Chack, Diswar, & Gupta, 2012). Common techniques applied to urban analysis fall into the general category of Spatial Hierarchical Clustering (SHC) frameworks (Carvalho, Albuquerque, Gilberto Rezende de Almeida, & Guimaraes, 2009). One SHC process is Ward's minimum variance method, which is a special case of the objective function approach originally presented by Ward (Ward, 1963). Ward proposes an agglomerative process that

uses features of polygons to cluster them based on similarity, optimizing "any [objective] function that reflects the investigator's purpose."

Subsequent authors have built on Ward's model to develop processes that are more sophisticated than the tool we demonstrate in this paper. Some authors use more-sophisticated adjacency and distance computations on polygon-based clustering routines, and some authors use multivariate clustering processes.

Deepti Joshi (2011) and D. Joshi, Samal, and Soh (2009) do both – using spatial and non-spatial attributes of polygons applied to a polygon dissimilarity function (PDF), to cluster polygons. Spatial attributes in their models include degrees of adjacency and distance, separation by water, and other features. A software tool developed by researchers at the Australian Urban Research Infrastructure Network (AURIN) uses a modified Ward's algorithm does multivariate clustering to identify economic concentrations (Bishop et al., 2016 forthcoming; J. Day, S. Sturup, Y. Chen, and M. Budahazy, 2014). Many other SHC polygon clustering processes use univariate data, as our tool does. Srucca (2005) generates economic clusters in Italy; Feser and Isserman (2005) derive spatial industry clusters of American counties; Jia and Jiang (2012) cluster micro-data such as intersections or parcels.

One benefit of an SHC-based clustering algorithm is that there is no need for a "seed" spatial unit to begin the clustering process. However, we cannot

guarantee that these clusters will center around the places that have meaning as urban cores, like capital cities or major population centers. There is advantage to simple models (Badland et al., 2013) that are usable by time-poor aid agencies and governments, and that allow the user to define the centers of important places.

Given the limitations of the different approaches, a new suite of tools is desirable if we are to arrive at a less-discretionary way to identify urban regions and rural areas that are concordant with Census geographies, capable of parsing large agglomerations into functional regions, usable anywhere in the world with only minimal and typical data as inputs, centered on the places we know are significant, assisted by an established process and software, and free for anyone to take and build upon. The process presented here extends Uchida and Nelson's (2008) Agglomeration Index framework, improving on some of the limitations of their and others' processes and adding to this necessary suite of tools.

The need for good processes to identify urban regions is not an abstract or academic one. We developed an early version of the process we describe here to define functional metropolitan areas for a national urbanization study of Indonesia (The World Bank, 2012) because metropolitan regions as officially defined were not suitable for a national study of cities and economic development. Unofficial but urbanized areas are often left out of major processes such as urban planning. For instance, in China, urbanized villages on the peri-urban fringes are often

entirely ignored in urban planning processes (Yang, Day, & Han, 2015). In Author Day's personal experience in peri-urban Port Vila after the Category 5-plus Tropical Cyclone Pam devastated the country in March 2015, peri-urban settlements were ignored for more than a month by government and aid agencies – at least partly because the extent of their urbanization was misunderstood. These processes are still being actively developed in advanced economies with extensive data on their urban areas, e.g., (OECD, 2012). Defining urban agglomerations in a simpler way is necessary for places that do not have such sophisticated data at the ready.

3 The Algorithm

We start with Uchida and Nelson's (2008) notion that an urban agglomeration can be built by starting with a seed with a certain population density, and then adding places to the agglomeration according to whether they meet density and travel-time criteria. Uchida and Nelson start with a list of human settlements and population data provided by the Center for International Earth Science Information Network (CIESIN) at Columbia University. Using various datasets on topology, waterways, and roads, they compile a raster *cost surface* with a one kilometer squared grid. This cost surface reflects the time it takes to travel through that grid, and allows for the computation of the path of least resistance

between places on the grid. They then overlay these parameters and compile metropolitan areas from the Global Rural-Urban Mapping Project (GRUMP) grid based on the following criteria:

- 1) A “seed” with more than a certain population, e.g., 20,000 or 50,000, provides the core of a metropolitan area. This seed is a grid square that meets the population requirements.
- 2) Grid squares are added to this metropolitan core if they meet travel time and density criteria, for example, grid squares that are reachable within 60 minutes of travel time and have population densities of more than 150 persons per square kilometer.

We have developed a logic that addresses some of the critical failures of Uchida and Nelson’s methods we describe above in the Background section. These include allocating contested border geographies and parsing contiguous metropolitan agglomerations into functional regions.

3.1 Algorithm Parameters

Keeping the core logical structure of Uchida and Nelson’s Agglomeration Index, we use travel time, population density, and seed population as our primary decision factors for the construction of urban agglomerations. Sections 3.2 and 3.33

describe the data that the software uses, the parameters that must be specified, and the decision structures it uses to produce the agglomerations.

Our algorithm parameters are summarized in Table 1, including the input parameters to construct urban agglomerations for Sri Lanka that are presented in Section 6. The foregoing sections describe these parameters in more detail. The foregoing sections also describe how these user inputs are used in the decision structure of the algorithm.

3.2 Seeds

We apply Uchida and Nelson's concept of "seeds" in this work, though we operationalise the seeds differently. Each agglomeration requires a seed, from which travel times are computed to neighbouring polygons and the parameters applied to test whether a polygon shall be included in the agglomeration. In the polygon-based algorithm, a seed is always comprised of an administrative polygon.

The software determines whether a polygon shall be designated as a seed. There are two criteria applied to a polygon to test whether it can be used as a seed. First, that seed must contain a city with a population above a certain threshold. The tool allows the user to define this threshold. The software asks for a file called, "Seed cities," as indicated in Table 1. These are point geographies indicating

the center of an urban settlement. This list serves as a set of *potential seeds*. The tool automatically excludes settlement points that have populations below the user-defined threshold.

After a potential seed is designated, there is one more step in defining metropolitan seeds. This process also serves to parse large contiguous regions into functional metropolitan areas. There may be multiple adjacent polygons that are designated as seeds even though these could belong to the same metropolitan area. To use each of these as a metropolitan seed would result in a single effective metropolitan region being erroneously divided into as multiple metropolitan regions by the algorithm. To address this problem, we allow seed polygons to be merged into a single seed if they meet either of two merging criteria: adjacency (they share a common border of any length) and distance (measured by polygon centroids being less than a user-specified distance apart, as-the-crow-flies). A default distance of 30 kilometers is built into the tool.

In this software, merging is not a literal process, where the boundaries and attributes of two polygons are combined. Our software maintains the integrity of the original polygons. When two seed polygons are merged in our tool, a notation is made in the data that assigns the *merged seed* to the *target seed* (the target seed is always the seed with the larger population). The tool marks the merged seed with an identification number that is identical to that of the target seed. Each

polygon retains its own original attributes, but the new attribute, i.e., identification number, is added that indicates that the two seeds are merged.

3.3 Allocation of contested border polygons

Once seeds are defined, the logic of the algorithm works by allowing each seed to “claim” those polygons that meet the decision travel-time and population-density criteria. This process is repeated for all seeds one by one, and can sometimes result in contested polygons claimed by more than one seed, particularly in the border regions of two adjacent metropolitan areas. In a case where two or more seeds claim a polygon, a decision algorithm is applied to assign the polygon to only one seed. This decision algorithm is based on a gravity measure with the following formula:

$$GRAVITY_i = \frac{POPULATION_j}{(DISTANCE_{ij})^2}$$

where $GRAVITY_i$ is the gravity measure for a given non-seed polygon, $POPULATION_j$ is the population of the seed polygon, $DISTANCE_{ij}$ is the Euclidean distance between polygon i and potential seed polygon j . The gravity measure is computed for each polygon, which is then assigned to the seed with the largest gravity value. This process occasionally results in agglomerations that are not contiguous. This can happen because the gravity measure is influenced by both

population and distance. We point out examples of non-contiguous regions in Section 4 below.

4 Tool Design and Data Structures

This section describes the elements of our software, including the user-defined functions and the decision structures that a user interested in running the source code should understand. This section also details the data that the algorithm is capable of taking as inputs. All of the processes described here are coded using Cran R and the source code and sample data are available at www.github.com/yiqunc/polygonaa.

4.1 Inputs

The user-provided data required to run our algorithm are minimal and are summarized in the first three rows of Table 1. They include a list of potential seeds, polygons, and population within each seed area and administrative polygon. Our tool provides polygon-based analysis that is generally compatible with Census geographies. Any administrative polygon data can be made compatible with the software. The tool automatically obtains travel data from online sources, which we describe below.

4.2 Seeds

The software determines whether a polygon shall be designated as a seed. There are two criteria applied to a polygon to test whether it can be used as a seed. First, that seed must contain a city with a population above a certain threshold. The tool allows the user to define this threshold. The software asks for a file called, “Seed cities,” as indicated in Table 1. These are point geographies indicating the center of an urban settlement. This list serves as a set of *potential seeds*. The tool automatically excludes settlement points that have populations below the user-defined threshold.

After a potential seed is designated, there is one more step in defining metropolitan seeds. There may be multiple adjacent polygons that are designated as seeds even though these could belong to the same metropolitan area. To use each of these as a metropolitan seed would result in a single effective metropolitan region being erroneously divided into as multiple metropolitan regions by the algorithm. To address this problem, we allow seed polygons to be merged into a single seed if they meet either of two merging criteria: adjacency (they share a common border of any length) and distance (measured by polygon centroids being less than a user-specified distance apart, as-the-crow-flies). A default distance of 30 kilometers is built into the tool, but users can specify another threshold.

In this software, merging is not a literal process, where the boundaries and attributes of two polygons are combined. Our software maintains the integrity of the original polygons. When two seed polygons are merged in our tool, a notation is made in the data that assigns the *merged seed* to the *target seed* (the target seed is always the seed with the larger population). The tool marks the merged seed with an identification number that is identical to that of the target seed. Each polygon retains its own original attributes, but the new attribute, i.e., identification number, is added that indicates that the two seeds are “merged.”

4.3 Travel times

Travel times between seeds and other polygons are computed from centroid to centroid. We assume that the traveler is moving in a rubber-wheeled vehicle on the road network. For each non-seed polygon, the tool calls the direction APIs described in Section 5 to compute travel time to its possible seed, which is the one has the largest gravity value as discussed in Section 3.3. The non-seed polygon will merge to that seed if the computed travel time is within the predefined threshold.

4.4 MAUP issue

We also realise and acknowledge that our current data preparation steps are impacted by MAUP (Modifiable Areal Unit Problem). Basically, any spatial analysis method containing point-to-polygon data aggregation steps can be challenged by

either of the two types of MAUP, which are scale effect and zone effect. When preparing the polygon data for our tool, we aggregate population density and minimum population from grid data to administrative boundaries; hence MAUP certainly impacts the outputs. While, in our case, the scale effect is more significant than the zone effect because the openly available population grid data has to be aggregated at the certain administrative boundary level rather than any arbitrary zones. To alleviate the scale effect in urban agglomeration, the key is to select a certain administrative boundary level that can best represent “seeds”. It is not necessarily to be the finest spatial unit, but an appropriate one that is suitable for that country. Local knowledge is definitely desired here to make this kind of decisions.

Anyway, MAUP belongs to data preparation issues, and it won't impact the tool's calculation logics. The tool itself should work with any polygon data regardless of how it is generated. For example, if population data is directly available at a certain administrative boundary level, users can directly work it using our tool and the outputs will be MAUP free.

5 Demonstration Data

The tool presented here takes advantage of free data that is generally available online from country websites and other open data hosts, including administrative

polygons and Census data as reported by country governments. In the demonstration we give in this paper, we use data that is available publically and free of charge. We use population and administrative data from Sri Lanka for 2011. We use current travel time because we do not have access to historic travel data, which could have the effect of producing larger metropolitan areas, since the road system has improved in the country over that time.

5.1 Polygons and population data

Table 1 describes the data we use in this paper. Many country governments make sub-national population data from national Censuses, available on their websites. The Database of Global Administrative Areas (GADM) publishes administrative maps for many countries at different sub-national administrative levels. The data are free and available online.

For this analysis, we use government-defined administrative polygons published on GADM. In Sri Lanka, the smallest administrative divisions available from GADM are Divisional Secretariat Divisions, or DS-Divisions. The country has 323 DS-Divisions. We use population data from the Sri Lankan Census website, also provided at the level of DS-Division.

5.2 Seeds

In our demonstration on Sri Lanka in Section 6, the administrative unit we use is the DS-Division. Thus, some of the DS-Divisions will be specified as seeds. Uchida and Nelson use 20,000, 50,000, and 100,000 as minimum seed population thresholds, so we follow this process in producing our empirical outcomes below.

In our analysis, we define cities according to designations published by the Government of Sri Lanka, though a user could choose to define cities in other ways. Table 2 lists the potential seed cities for Sri Lanka. In Sri Lanka, the government provides three levels of urban designation: Urban Councils (UCs), Municipal Councils (MCs), and Pradeshiya Sabhas (PSs), which Table 2 also indicates.

After a potential seed (in the case of Sri Lanka, a DS-Division) is designated, there is one more step in defining metropolitan seeds. There may be multiple adjacent polygons that are designated as seeds even though these could belong to the same metropolitan area. The region around Colombo, for instance, has four cities with more than 100,000 people and eight cities with populations in excess of 50,000. These are spread over adjacent DS-Divisions. To use each of these as a metropolitan seed would result in a single effective metropolitan region (e.g., Colombo) being erroneously divided into as multiple metropolitan regions by the algorithm. To address this problem, we allow seed polygons to be merged into a

single seed if they meet either of the two merging criteria. A default distance of 30 kilometers was used for the agglomerations generated here.

Here we need to introduce a tuple (a, b, c) to concisely describe the parameter configurations. The first number in tuple stands for population density (persons per square kilometer); the second number is the minimum population (number of persons), and the third number refers to travel time in minutes. We will keep using these tuples in the rest of this paper.

Figure 1 illustrates how the merging process works in the region around Colombo. In the (150; 50,000; 60) configuration, the region around Colombo has seven cities that are candidate seeds. These are shown in Figure 1a. When our seed-merging logic is applied, all seven of these seeds collapse to a single seed, centered at Colombo. Figure 1b shows the new seed configuration after the seed merging process. Figure 1 also shows the resulting metropolitan agglomerations around Colombo if the seeds are not merged (Figure 1a) and if they are merged (Figure 1b).

[Insert Figure 1]

In this software, merging is not a literal process, where the boundaries and attributes of two polygons are combined. Our software maintains the integrity of the original polygons. When two seed polygons are merged in our tool, a notation is made in the data that assigns the *merged seed* to the *target seed* (the target seed

is always the seed with the larger population). The tool marks the merged seed with an identification number that is identical to that of the target seed. Each polygon retains its own original attributes, but the new attribute, i.e., identification number, is added that indicates that the two seeds are merged.

5.3 Travel times

Travel times in this demonstration are computed using two free and widely-available application programming interfaces (APIs): Mapquest Direction API (Mapquest, 2014) and Google Direction API (Google, 2014). This approach is an improvement over Uchida and Nelson's use of the GRUMP data in that it is free, generally up-to-date, and considers traffic and road construction in many places. Our tool automatically sends requests to these APIs to compute travel times, so no additional steps or expertise are required by the user to obtain travel data.

The Mapquest API is called first by the tool because it allows unlimited point-to-point travel time queries for free. The software falls back to Google API when Mapquest fails to generate travel time. The Google API restricts free requests to 2,500 per day and requires payment when extra requests occur. In the small number of instances where neither Mapquest nor Google can generate travel time, the mean travel speed for all computed point-to-point segments is applied to the missing segment to arrive at an estimated travel time. For each non-seed

polygon, the algorithm calls the API to compute travel time to its possible seed, which is the one has the largest gravity value as discussed in Section 3.3.

6 Algorithm Performance, Sri Lanka

This section describes the application of our Agglomeration Polygon Algorithm to Sri Lanka. We compare the outputs of the algorithm under different combinations of parameter specifications.

6.1 Algorithm estimations of national urbanization rates

Table 3 shows the computed urbanization rates for Sri Lanka for 2011. Depending on the parameters applied to our algorithm, 2011 urbanization rates vary between 42 and 82 percent. We note that these figures do not agree with United Nations urbanization estimates. The *World Urbanization Prospects* (WUP) dataset (UNESA, 2011) indicates a 2011 urban population of 3,137,000 for the country, and a total national population of 20,263,723. This results in an urbanization rate of 15 percent. The WUP uses country-defined urban and rural designations.

Our algorithm-computed urbanization rates for Sri Lanka consistently show urbanization rates that are significantly higher than published those published by the WUP urbanization statistics. Figures 2 and 3 illustrate national urbanization rates computed using the algorithm, for six combinations of Agglomeration

Polygon Algorithm parameters. Figure 2 shows our computed urbanization rates based on population, i.e., the proportion of the population estimated to be living in urban areas, compared with the WUP rates. Figure 3 gives urbanization by land area, i.e., the proportion of land area estimated to be within metropolitan regions. The WUP does not publish urbanized land areas so that metric is not included in the figure. For figures, travel time and seed population thresholds are allowed to vary, and the population density threshold is held constant at 150 persons per square kilometer.

[insert Figure 2.]

[insert Figure 3.]

6.2 Number and size of agglomerations

Figures 4 and 5 show the distribution of population and land area contained in the metropolitan regions under the six parameter combinations. These illustrate a number of properties of the algorithm. First, the number of agglomerations identified is inversely related to the size threshold for the seed cities. That is, smaller seed population thresholds tend to result in more identified agglomerations. This is expected. The algorithm only allows seed merging if a distance or adjacency condition is met, so small seeds that are not sufficiently near other seeds can retain their independence.

[insert Figure 4.]

[insert Figure 5.]

Under the (150; 100,000; 60) and (150; 100,000; 90) configurations, the algorithm only identifies four metropolitan regions, which from Table 3 contain 49 and 57 percent of the national population in aggregate. In Table 3, under the most restrictive parameter combinations reported here (500; 100,000; 60), the number of metropolitan agglomerations remains at 4, while these large agglomerations only hold 41.85 percent of the national population. A graph of these results is not shown here for brevity.

Second, the population and land area classified as rural decreases (i.e., urban-classified populations increase) as the seed city thresholds become more restrictive. Similarly, fewer people and places are classified as rural as travel time thresholds increase. Also, the size of some of the metropolitan regions fluctuates dependent on the parameter specification. This is most obvious in the larger regions such as Colombo and Kandy. Although Colombo remains a primate city under all specifications, its population share varies from 26 to 37 percent, and Kandy's from 10 to 16 percent of the national population. Although it is not apparent from Figures 6 and 7, the different parameter specifications allocate different populations to different metropolitan areas in some cases, as we will see in Section 6.3.

6.3 Configuration of Agglomerations

Figure 6, 7 and 8 show maps of the resulting metropolitan regions under three parameter configurations, (150; 20,000; 60), (150; 50,000; 60) and (150; 100,000; 60). Table 4, 5, and 6 enumerate the populations and land areas under the three configurations. Under the parameter combinations in Figure 7 (150; 50,000; 60), there are 8 metropolitan agglomerations and the population urbanization rate is 63 percent of the national population. Under the parameter combinations in Figure 8 (150; 100,000; 60), the number of metropolitan agglomerations drops to 4, and the urbanization rate drops to 49 percent. As a naming convention, we use the name of the largest city in the agglomeration.

[insert Figure 6]

[insert Figure 7]

[insert Figure 8]

As in Section 6.2, the spatial mapping of the computed agglomerations tells us about the workings of the algorithm. First, smaller seed population thresholds generally lead to more agglomerations being identified. Comparing Figures 6 and 7 to Figure 8, we can see that some cities with smaller populations, e.g., Trincomalee and Jaffna, are able to form the kernel of an agglomeration when the seed threshold is allowed to be smaller. Users, then, should be aware of the types of areas they wish to be included as metropolitan when they are setting the

algorithm parameters. From Table 2, we can see that Jaffna has a population of just over 93,000 people – qualitatively near to the 100,000 population threshold required in the (150; 100,000; 60) configuration. This illustrates an important point about quantitative processes like the one described in this paper: algorithms are unforgiving.

It is the purview and responsibility of the analyst to pay attention to the local context in order to make good decisions about setting parameter configurations. When we showed the outputs of the tool to policy makers in Sri Lanka, their local knowledge suggested that using a population threshold of 20,000 or 100,000 produced agglomerations that were not aligned with their understanding of their cities. In the dense Southwest of the country, the (150; 20,000; 60) parameter specification identifies too many agglomerations, and the (150; 100,000; 60) configuration too few. The (150; 50,000; 60) specification, however, looks just about right. This suggests that in a country or region like Sri Lanka, Java, Indonesia, or the Pearl River Delta in China, where urbanization is more continuous but of lower density, parameter specifications will be different than in regions like sub-Saharan Africa or Eastern Europe, where urbanization patterns are different. Where urban areas are discontinuous and large, e.g., the east coast cities of China, perhaps larger seed sizes would suffice. The configuration of the algorithm parameters, then, is a matter of local knowledge and

judgment. There is also room for the empirical fitting of the algorithm based on local conditions. We take this up in the conclusion.

The second insight the maps offer about the workings of the algorithm, is that some configurations result in metropolitan regions that are not spatially contiguous. This can be seen in Figures 6 and 7, with the Vavuniya, where a northern core DS-Division and two additional DS-Divisions to the south are separated by rural DS-Divisions. This discontinuity occurs because the southern DS-Divisions in the agglomeration are sufficiently urbanized to meet density standards, sufficiently close to meet travel-time standards, but lack a suitably-sized urban core of their own to define an independent agglomeration.

This highlights the need for analysts using our tool to understand the local context. This tool is not meant to replace a thoughtful understanding of regional geography and structure; it is simply meant to supplement the knowledge that analysts and local policy makers already have about a place. For a simple tool like this, the possible confounding factors are many. For instance, some government metropolitan designations include watershed and forest areas that would have insufficient population to be classified as urban; others have geographic features such as mountains and lakes contained within them, which reduces their population densities.

A third insight that the maps offer is that some agglomeration configurations may not always reflect the relative local importance of cities and towns. The Vavuniya agglomeration in Figures 6 and 7 is so named because Vavuniya is the largest city, by population, in the set of assembled polygons. However, someone with local knowledge would be able to identify a problem with this agglomeration configuration; namely, that the dominant city in that agglomeration is the fast-growing Urban Council of Anuradhapura. Table 2 offers some insight into why the misnaming happens in this case: Vavuniya has a larger population than Anuradhapura (72,000 versus 61,000), so it is selected by the algorithm as the seed.

Local knowledge, then, can help an analyst to overcome the limitations inherent in any automated process such as this one. Based on the aim of the study being undertaken, an analyst with local knowledge could choose to adjust the metropolitan boundaries to include the interstitial DS-Divisions, exclude the disjointed southern DS-Divisions, rename the agglomeration appropriately, or make other changes in the algorithm's output to reflect local context. At very least, the disjointed Vavuniya metropolitan region suggests that an adjustment to the algorithm's outputs is worth exploring.

Table 3 illustrates that, without exception, smaller seed thresholds will result in a higher proportion of the national population being classified as

belonging to a metropolitan area. One way that this works is that smaller seeds form metropolitan regions where none existed with larger thresholds, as we discuss above, e.g., Jaffna. There is, however, another process by which the number of metropolitan agglomerations increases with smaller seed thresholds. Namely, a region that is classified as a single large metropolitan agglomeration in a (150; 50,000; 60) configuration will be split into multiple smaller agglomerations in a (150; 20,000; 60) configuration. Compared with the large singular Colombo region presented in Figures 7, the Puttalam region is detached from Colombo and forms a separate agglomeration in Figure 6.

Although large regions splitting into smaller regions is generally a process associated with smaller seed thresholds, it is not exclusive to them. In Figure 7, under the (150; 50,000; 60) configuration, Colombo is classified as one large, contiguous region. In Figure 8, with the seed threshold increased to 100,000 but both other variables remaining the same, the Colombo region is split into two smaller regions, one centered around the city of Colombo and one around the city of Gampaha.

The fragmentation of the Colombo region under the larger seed threshold occurs because of the seed-merging decision structure of our algorithm. Figure 9 shows the three largest seeds in proximity to Colombo, along with their populations (the region has four other cities with populations greater than 50,000,

but we leave those out of this graphic for simplicity). If we choose 50,000 as the minimum seed population, all three areas will be eventually merged into Colombo by the algorithm. In this case, Gampaha is merged with Colombo first, and then serves as a 'bridge' seed, connecting Negombo and Colombo.

[insert Figure 9.]

If we choose 100,000 as the minimum seed population, only Negombo and Colombo will remain as viable seeds (Gampaha does not meet the population threshold requirement). However because they are 30.2 kilometers apart, Negombo cannot be merged to Colombo (our distance threshold for merging seeds is 30km). Thus, both Gampaha and Colombo remain separate agglomerations.

Finally, the maps presented here tell us something about how the algorithm allocates border polygons sitting at the edges of metropolitan agglomerations. Between Figures 7 and 8, a number of DS-Divisions sitting at the borders of the various metropolitan regions shift from urban to rural, rural to urban, or from belonging to one region or another, for instance, between Colombo and Kandy. The stark differences between metropolitan areas under different parameter configurations – including their size, boundaries, and contiguity – is worth discussing at length. We take this up in the conclusion.

6.4. Comparison and Notes on Parameter Validation

One general issue with defining the extents of urban agglomerations is that is difficult to assess the accuracy of the estimations. This is a double-edged sword: there is a need for a tool like this because the extents of metropolitan areas are difficult to effectively measure, but on the other hand, there are no objective measures against which we can validate the extents that we compute. In many developed countries, metropolitan extents are defined by the census bureaus, for instance, the United States Census Bureau defines urban extents based on commute behavior to metropolitan cores. In these places, a selection of parameter configurations for our algorithm that matches the resulting agglomerations to government designations, would be relatively easy to settle upon. However, it is not these places where the designations are most needed. It is developing countries, where accurate urban/rural designations are perhaps more important, that such data for validation is most lacking. This speaks to a need for our tool, but again, does not provide means for validating it in developing-country contexts.

As we have already discussed, government metropolitan designations are influenced by political issues and can ignore urbanization that has spread into districts still officially designated as rural. With a tool that shows such broad variation in the urbanization estimates arising from different parameter

specifications, it is a difficult task to decide on an appropriate parameter configuration, and which set of outputs represents the true urbanization pattern.

For now, we will need to rely on local knowledge for this type of validation. Forthcoming data from the CIESIN may provide a path toward a validation process. Currently, the CIESIN website contains images of urban extents for 1995. These urban extents are based on computations that take into account night-time lights observed from satellite images, population characteristics from national Censuses, and other relevant demographic information. CIESIN's website indicates that updated Night-time Lights data may soon be forthcoming (CIESIN, 2013). These data are not independent from political considerations, but together with our algorithm computations, could give rise to a process for cross-validation and comparison.

It is useful to somehow place our findings in context. The WUP provides a summary of urban population distribution categorized by urban-size designations (UNESA, 2011). However, the WUP data only catalogs agglomerations with populations greater than 500,000, so we cannot compare the data for agglomerations smaller than a half-million people. Also, the WUP data is derived from country government reporting, which suffers from the problems we identify above (urbanized areas still classified as rural, etc.). Still, the comparison for larger classes of cities provides context for interpreting the algorithm estimates.

Table 7 below gives the number of agglomerations, urban population, and percent of the urban population held in each of five city size classifications and overall. Urbanization estimates are given using our Agglomeration Polygon Algorithm and the WUP data. The table results based on the (150; 50,000; 60) configuration and the (150; 100,000; 60) configuration.

Both tables indicate that our tool under these configurations places the vast majority of the urban population in larger metropolitan regions. Table 7 indicates that, under these configurations, there are three resulting agglomerations containing one to five million people, containing between 80 and 93 percent of the country's urban population.

The pattern in the WUP is reversed. WUP reports that 78 percent of the urban population lived in small cities of less than 500,000 in 2010, and the one primate city (presumably Colombo) contained 22 percent of the urban population, or 687,000 people. Our algorithm's overall urban population estimates are also much larger than the WUP estimates, as shown in Table 7. WUP estimates that 3.1 million Sri Lankans lived in urban regions in 2010; our algorithm places that number at triple and quadruple the WUP estimates, depending on the configuration of the algorithm.

The WUP statistics do not align with McGee's (McGee, 1969, 1995) assertions that much of South Asia is low-density urbanization. These

misalignments point to a need to further investigate the problem. We provide this tool as a starting point for such an inquiry.

7 Conclusions

We provide this open source tool using, as a starting point toward a more robust model of identifying urban agglomerations. We hope that future tools and approaches will refine our methods to build a model of identifying urban agglomerations that is less discretionary – that is, a model that internalizes more of the complexities that are inherent in defining urban regions.

The disparity between our model’s outputs and the urbanization estimates from the United Nations should alarm the skeptic of our process. If we take the WUP data as a baseline, then our algorithm has failed in Sri Lanka, producing far higher estimates of urbanization than established methods do. Indeed, we have likely overestimated the urban population of Sri Lanka, due to what Isserman (2005) calls the “county trap” – the misassignment of population due to the large and urban-rural heterogeneous geographies used in the analysis (counties are large geographic units in the USA, where Isserman’s work is based). This can be at least partly remedied by using the polygon tool in conjunction with the point-based algorithm (Day *et al.*, 2016) to obtain estimates of the overcounts, or combining the tool outputs with remotely-sensed data such as nighttime lights.

We see a number of places where future work can refine our method, which is why we open source the tool.

If, however, we remember that those established methods (WUP estimates from country-defined urban regions) are not grounded in any systematic method that is applicable in all settings, and if we further remember that countries have various political and historical motivations for defining their urban regions the way that they do, then we must acknowledge that a systematic process for identifying urban agglomerations could improve the way we currently identify places as urban. Our process uses spatial demographic data to identify urban agglomerations empirically, thus removing some of the political and historical barriers to urban designation. Systematic international urban definitions clear the way for rigorous cross-country analyzes and coherent regional urban policy. It also provides a tool for analysts to use to understand whether there are significant urbanized areas that are not acknowledged as such due to political and historic concerns, or simply because urbanization happens faster sometimes than governments can reclassify rural areas to urban ones. It follows that, if countries are not even officially acknowledging their urban populations, they might also not be effectively planning to accommodate them. This and future tools can help to change these realities.

The tool's inputs raise other issues, such as the comparability of administrative divisions across countries. For instance, India's Level 2 administrative polygons (a GADM designation) are larger in area and have larger populations than those of Sri Lanka. Sri Lanka in 2012 had 323 Level 2 polygons with an average of 62,277 people and a land area of 208 square kilometers. India's 2011 GADM data has 593 Level 2 polygons with an average of 1.75 million people and a land area of 5,315 square kilometers. Agglomerations constructed from administrative polygons with different spatial configurations or populations may have compatibility issues that need to be worked out by analysts according to their purposes. We do not include a sensitivity analysis of the impact of coarser or finer spatial units because of space constraints, and also because this sensitivity will be different in every country. However, we do alert our readers to pay attention to such considerations when using our tool. Our tool allows analysts to use different thresholds and parameters for different settings if appropriate, and issues like these are why we open source our code to allow anyone to improve it.

References

- Badland, H., M. White, G. MacAulay, S. Eagleson, S. Mavoa, C. Pettit, & B. Giles-Corti. (2013). Using Simple Agent-Based Modeling to Inform and Enhance Neighborhood Walkability. *International journal of health geographics*, 12(1), 1.
- Bagan, H., & Y. Yamagata. (2014). Land-Cover Change Analysis in 50 Global Cities by Using a Combination of Landsat Data and Analysis of Grid Cells. *Environmental Research Letters*, 9(6), 064015 (064013 pages).
- Bishop, I. D., C. Pettit, S. Eagleson, A. Rajabifard, B. H., J. Day, J. Furler, M. Kalantari, S. Sturup, & M. White. (2016 forthcoming). Using an Online Data Portal and Prototype Analysis Tools in an Investigation of Spatial Liveability Planning. *International Journal of E-Planning Research*.
- Carvalho, A. X. Y., P. H. M. Albuquerque, J. Gilberto Rezende de Almeida, & R. D. Guimaraes. (2009). Spatial Heirarchical Clustering. *Revista Brasileira de Biometria*, 27(3), 411-442.
- CIESIN. (2013). Night-Time Lights Can Help Illuminate Trends in Urbanization. Retrieved from <http://blogs.ei.columbia.edu/2013/11/11/night-time-lights-illuminate-trends-in-urbanization/>
- Day, J., Y. Chen, P. Ellis, & M. Roberts. (2016). A Free, Open Source Tool for Identifying Urban Agglomerations Using Point Data. *Spatial Economic Analysis*, 11(1), 67-91.
- Day, J., & P. Ellis. (2013a). Growth in Indonesia's Manufacturing Sectors: Urbanization or Localization Economies? *Regional Science Policy and Practice*, 5(3), 343-368.
- Day, J., & P. Ellis. (2013b). Urbanization for Everyone: The Benefits of Urbanization in Indonesia's Rural Regions. *Journal of Urban Planning and Development*, 140(3), 04014006-04014001 - 04014006-04014009.
- Day, J., & B. Lewis. (2013). Beyond Univariate Measurement of Spatial Autocorrelation: Disaggregated Spillover Effects for Indonesia. *Annals of GIS*, 19(3), 169-185.
- Day, J., S. Sturup, Y. Chen, and M. Budahazy. (2014). *An Open-Source Tool for Identifying Industry Clusters: A Demonstration in the Northwest Corridor of Melbourne, Australia*. Paper presented at the Australian Regional Development Conference, Albury.
- Doll, C. N. (2008). Ciesin Thematic Guide to Night-Time Light Remote Sensing and Its Applications. *Center for International Earth Science Information Network of Columbia University, Palisades, NY*.

- Feser, E., & A. Isserman. (2005). *Clusters and Rural Economies in Economic and Geographic Space, Working Paper*.
- Gamba, P., & M. Herold (Eds.). (2009). *Global Mapping of Human Settlement: Experiences, Datasets, and Prospects*. Boca Ratan: CRC Press.
- Google. (2014). Google Maps Api Web Services Specs. Retrieved from <https://developers.google.com/maps/documentation/directions/>.
- Government of Sri Lanka. (Date not declared). *Declaration of Urban Areas and Approved Development Plans*.
- Jia, T., & B. Jiang. (2012). *Scaling Property of Urban Systems Using an Entropy-Based Hierarchical Clustering Method*. Paper presented at the Multidisciplinary Research on Geographical Information in Europe and Beyond. Proceedings of the AGILE'2012 International Conference on Geographic Information Science, Avignon.
- Joshi, D. (2011). *Polygonal Spatial Clustering*. (PhD), University of Nebraska - Lincoln.
- Joshi, D., A. Samal, & L.-K. Soh. (2009). *A Dissimilarity Function for Clustering Geospatial Polygons*. Paper presented at the 17th International Conference on Advances in Geographic Information.
- Lo, C. P., D. A. Quattrochi, & J. C. Luvall. (1997). Application of High-Resolution Thermal Infrared Remote Sensing and Gis to Assess the Urban Heat Island Effect *International Journal of Remote Sensing*, 18(2), 287-304.
- Mapquest. (2014). Open Directions Service Developer's Guide. Retrieved from <http://open.mapquestapi.com/directions/>.
- McGee, T. G. (1969). Urbanization or Kotadesasi? Evolving Patterns of Urbanization in Asia" In F. J. Costa, A. K. Dutt, L. J. C. Ma, & A. G. Noble (Eds.), *Urbanization in Asia: Spatial Dimensions and Policy Issues*. Honolulu: University of Hawaii Press.
- McGee, T. G. (1995). Metrofitting the Emerging Mega-Urban Regions of Asean: An Overview. In T. G. McGee & I. M. Robinson (Eds.), *The Mega-Urban Regions of Southeast Asia* (pp. 3-26). Vancouver: UBC Press.
- Møller-Jensen, L. (2013). Methods for Texture-Based Classification of Urban Fringe Areas from Medium and High Resolution Satellite Imagery. In J. R. Weeks, A. G. Hill, & J. Stoler (Eds.), *Spatial Inequalities: Health, Poverty, and Place in Accra, Ghana*. Dordrecht, Heidelberg, New York, London: Springer.
- Morrill, R., J. Cromartie, & G. Hart. (1999). Metropolitan, Urban, and Rural Commuting Areas: Toward a Better Depiction of the United States Settlement System. *Urban Geography*, 20(8), 727-748.
- OECD. (2012). *Redefining Urban: A New Way to Measure Metropolitan Areas; Functional Urban Areas in Oecd Countries*. Retrieved from

- Pesaresi, M., G. Huadong, X. Blaes, D. Ehrlich, S. Ferri, L. Gueguen, M. Halkia, M. Kauffmann, T. Kemper, L. Lu, & M. A. Marin-Herrera. (2013). A Global Human Settlement Layer from Optical Hr/Vhr Rs Data: Concept and First Results. *Selected Topics in Applied Earth Observations and Remote Sensing, IEEE Journal*, 6(5), 2102-2131.
- Rotgé, V. (2001). *Rural-Urban Integration in Java: Consequences for Regional Development and Employment* Aldershot, Ashgate Publishing.
- Scott, J. C. (1998). *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven: Yale University Press.
- Shim, D. (2014). Remote Sensing Place: Satellite Images as Visual Spatial Imaginaries. *Geoforum*, 51, 1520160.
- Srucca, L. (2005). *Clustering Multivariate Spatial Data Based on Local Measures of Spatial Autocorrelation. An Application to the Labour Market of Umbria*. Research Division, Federal Reserve Bank of St. Louis. Ideas. St. Louis. Retrieved from <http://ideas.repec.org/p/pia/wpaper/20-2005.html>
- Tatem, A. J., A. M. Noor, & S. I. Hay. (2005). Assessing the Accuracy of Satellite Derived Global and National Urban Maps in Kenya. *Remote sensing of environment*, 96(1), 87-97.
- The World Bank. (2012). *The Rise of Metropolitan Regions: Towards Inclusive and Sustainable Regional Development*, World Bank Report, 71740. Retrieved from <http://documents.worldbank.org/curated/en/2012/08/16587797/indonesia-rise-metropolitan-regions-towards-inclusive-sustainable-regional-development>
- UCHIDA, H., & A. NELSON. (2008). *Agglomeration Index: Towards a New Measure of Urban Concentration*. Retrieved from Washington, D.C.:
- UNESA. (2011). File 17a: Urban Population (Thousands), Number of Cities and Percentage of Urban Population by Size Class of Urban Settlement, Major Area, Region and Country, 1950-2025. Retrieved 08 February 2014, from United Nations Department of Economic and Social Affairs, Population Division (<http://esa.un.org/unup/CD-ROM/Urban-Agglomerations.htm>)
- Verma, M., M. Srivastava, N. Chack, A. K. Diswar, & N. Gupta. (2012). A Comparative Study of Various Clustering Algorithms in Data Mining. *International Journal of Engineering Research and Applications (IJERA)*, 2(3), 1379-1384.
- Wang, L., C. Li, Q. Ying, X. Cheng, X. Wang, & X. Li. (2012). China's Urban Expansion from 1990 to 2010 Determined with Satellite Remote Sensing. *Chinese Science Bulletin*, 57(22), 2802-2812.
- Ward, J. H. J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, 58, 236-244.

- Yang, X., J. Day, & S. S. Han. (2015). Urban Peripheries as Growth and Conflict Spaces: The Development of New Towns in China In T.-C. Wong & S. S. Han (Eds.), *Population Mobility, Urban Planning and Management in China: An Introduction* (pp. 95-112). Cham: Springer.
- Zha, Y., J. Gao, & S. Ni. (2003). Use of Normalized Difference Built-up Index in Automatically Mapping Urban Areas from Tm Imagery. *International Journal of Remote Sensing*, 24(3), 583-594.

Table 1. Tool Parameters and Data Inputs

Parameters	Description	For Sri Lanka
Administrative polygons	User can use polygons at any administrative scale	DS-Divisions
Population of administrative polygons	The population living within the administrative polygon	Population of DS-Divisions
Seed cities	User can designate any of the administrative polygons as seeds by inputting a list of seed areas	Cities within DS-Divisions
Seed merging distance threshold	User can define the road travel distance between polygon centroids, within which seeds can be merged	30km
Seed population threshold	User can specify the minimum population at which an urban concentration is considered to be a seed	20,000; 50,000; 100,000
Population density threshold	User can specify the minimum population density (ppl. per square kilometer) at which a polygon will be considered to be part of the metropolitan region	150; 300; 500
Travel time threshold	User can specify the maximum travel time for which a polygon will be considered to be part of a metropolitan region	30; 60; 90

Table 2. Seed Cities by Population, 2011

City	Municipal or Urban Council	Population Rank	Population
Colombo	MC	1	671,407
Dehiwala-Mount Lavinia	MC	2	218,455
Moratuwa	MC	3	184,233
Negombo	MC	4	135,765
Sri Jayawardenepura (Kotte)	MC	5	120,737
Kandy	MC	6	117,227
Kalmunai (incl. Sainthamarathu)	MC	7	103,074
Batticaloa	MC	8	101,419
Galle	MC	9	96,588
Jaffna	MC	10	93,854
Katunayake (Seeduwa-Katunayake)	UC	11	81,789
Matara	MC	12	81,235
Vavuniya	UC	13	72,169
Trincomalee	UC	14	66,507
Gampaha	MC	15	64,101
Anuradhapura	MC	16	61,242
Kolonnawa	UC	17	57,420
Ratnapura	MC	18	49,186
Puttalam	UC	19	44,811
Kattankudy	Not specified	20	43,881
Badulla	MC	21	42,733
Kalutara	UC	22	42,696
Matale	MC	23	40,047
Kinniya	UC	24	38,660
Panadura	UC	25	38,114
Beruwala	UC	26	37,682
Ja-Ela	UC	27	34,482
Peliyagoda	UC	28	33,333
Wattala-Mabola	UC	29	32,186
Kurunegala	MC	30	31,381
Gampola	UC	31	26,034
Chilaw	UC	32	25,865
Nuwara Eliya	MC	33	25,515
Weligama	UC	34	23,206
Avissawella	Not specified	35	22,408

Ambalangoda	UC	36	21,100
Ampara	UC	37	19,570
Kilinochchi	Not specified	38	19,024
Kegalle (Kegalla)	UC	39	18,293
Chavakachcheri	UC	40	14,780
Mannar	UC	41	13,109
Hambantota	Not specified	42	12,615
Point Pedro	UC	43	11,512
Moneragala	Not specified	44	7,666
Valvettithurai	UC	45	6,318

Source: Government of Sri Lanka Declaration of Urban Areas (Government of Sri Lanka, Date not declared)

Table 3. Estimated Urbanization Rates

Agglomeration Polygon Algorithm Parameters			2011 Agglomeration Polygon Algorithm Output		
Population Density (ppl per square km; min)	Seed Population (minimum)	Travel Time (minutes; maximum)	2011 Computed Percent Urban	2011 Census Population	2011 Computed Urban Population
150	20,000	60	74.74	20,263,723	15,145,954
150	20,000	90	82.21		16,658,153
150	50,000	60	62.61		12,687,776
150	50,000	90	74.25		15,046,721
150	100,000	60	49.06		9,940,866
150	100,000	90	57.39		11,628,692
300	20,000	60	67.34		13,645,318
300	20,000	90	72.74		14,739,438
300	50,000	60	59.56		12,070,079
300	50,000	90	66.21		13,415,777
300	100,000	60	47.53		9,631,906
300	100,000	90	52.8		10,699,748
500	20,000	60	56.97		11,545,198
500	20,000	90	58.44		11,841,804
500	50,000	60	51.15		10,365,258
500	50,000	90	53.1		10,759,724
500	100,000	60	41.85		8,480,575
500	100,000	90	44.35		8,987,367
Summary Statistics					
Mean			59.57		12,071,686
Minimum			41.85		8,480,575
Maximum			82.21		16,658,153

Table 4. Sri Lanka Metropolitan Agglomerations in the (150; 20,000; 60) Configuration, 2011

Agglomeration	Population 2011	Area (sq. km)
Colombo	6,530,040	5,184
Kandy	2,066,163	3,417
Matara Four Gravets	1,901,799	3,198
Kurunegala	870,571	2,413
Kalmunai	686,841	1,055
Ratnapura	549,131	1,397
Jaffna	509,888	848
Nuwara Eliya	411,839	909
Chilaw	403,491	1,127
Badulla	394,465	1,218
Puttalam	381,759	1,173
Vavuniya	221,903	1,065
Trincomalee Town and Gravets	218,064	470
Urban Total	15,145,954	23,474
Rural	5,117,769	42,136
Sri Lanka Total	20,263,723	65,610

Table 5. Sri Lanka Metropolitan Agglomerations in the (150; 50,000; 60) Configuration, 2011

Agglomeration	Population 2011	Area (sq. km)
Colombo	6,611,997	5,357
Kandy	2,470,077	3,580
Matara Four Gravets	1,042,637	2,021
Galle Four Gravets	926,369	1,235
Kalmunai	686,841	1,055
Jaffna	509,888	848
Vavuniya	221,903	1,065
Trincomalee Town and Gravets	218,064	470
Urban Total	12,687,776	15,632
Rural	7,575,947	49,978
Sri Lanka Total	20,263,723	65,610

Table 6. Sri Lanka Metropolitan Agglomerations in the (150; 100,000; 60) Configuration, 2011

Agglomeration	Population 2011	Area (sq. km)
Colombo	5,279,075	3,538
Kandy	2,544,019	3,719
Negombo	1,430,931	2,042
Kalmunai	686,841	1,055
Urban Total	9,940,866	10,353
Rural	10,322,857	55,257
Sri Lanka Total	20,263,723	65,610

Table 7. Sri Lanka Comparison of Computed Urbanization Characteristics, 2011, with World Urbanization Prospects Urbanization Characteristics 2010

City Size	Parameter	WUP	Algorithm (150; 50,000; 60)	Algorithm (150; 100,000; 60)
10 million or more	Number of agglomerations	0	0	0
	Percentage of urban population	0	0	0
	Population	0	0	0
5 to 10 million	Number of agglomerations	0	0	0
	Percentage of urban population	0	0	0
	Population	0	0	0
1 to 5 million	Number of agglomerations	0	3	3
	Percentage of urban population	0	79.8	93.1
	Population	0	10,124,711	9,254,025
500 000 to 1 million	Number of agglomerations	1	3	1
	Percentage of urban population	22.0	16.7	6.9
	Population	687,000	2,123,098	686,841
Less than 500,000	Number of agglomerations	.	2	0
	Percentage of urban population	78.0	3.5	0.0
	Population	2,450,000	439,967	0
TOTAL	Number of agglomerations	.	8	4
	Percentage of urban population	100.0	100.0	100.0
	Population	3,137,000	12,687,776	9,940,866

(.) denotes data not available or not computable

(150 : 20,000 : 60)

(150 : 20,000 : 90)

(150 : 50,000 : 60)

(150 : 50,000 : 90)

(150 : 100,000 : 60)

(150 : 100,000 : 90)

(150 : 20,000 : 60)

(150 : 20,000 : 90)

(150 : 50,000 : 60)

(150 : 50,000 : 90)

(150 : 100,000 : 60)

(150 : 100,000 : 90)

Minerva Access is the Institutional Repository of The University of Melbourne

Author/s:

Day, J;Chen, Y;Ellis, P;Roberts, M

Title:

A free, open-source tool for identifying urban agglomerations using polygon data

Date:

2017-03-01

Citation:

Day, J., Chen, Y., Ellis, P. & Roberts, M. (2017). A free, open-source tool for identifying urban agglomerations using polygon data. *Environment Systems and Decisions*, 37 (1), pp.68-87. <https://doi.org/10.1007/s10669-017-9623-z>.

Persistent Link:

<http://hdl.handle.net/11343/283314>