Routledge
Taylor & Francis Group

Spatial Economic Analysis

ISSN: 1742-1772 (Print) 1742-1780 (Online) Journal homepage: http://www.tandfonline.com/loi/rsea20

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

Jennifer Day, Yiqun Chen, Peter Ellis & Mark Roberts

Spatial Economic Analysis, 2016
Vol. 11, No. 1, 67–91, http://dx.doi.org/10.1080/17421772.2016.1102957

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

JENNIFER DAY, YIQUN CHEN, PETER ELLIS & MARK ROBERTS

(*Received August 2014; accepted August 2015*)

Abstract *This paper describes a software tool for identifying urban agglomerations in low-information settings. The framework outlined in this paper is designed to work using point data. Our tool and all required data are provided free and in open-source format. This paper describes the advantages and disadvantages of using point-based geographies in regional analysis, discusses the practical and ethical challenges of distinguishing urban from rural regions, details the function of our software, and directs the interested reader to the source code. The paper also examines the tool's outputs for Sri Lanka and compares them with published United Nations urbanization figures. Our outputs indicate that Sri Lanka's urban population is significantly undercounted in official statistics.*

Un outil *Open Source* gratuit pour l'identification d'agglomérations urbaines à l'aide de données ponctuelles

RÉSUMÉ La présente communication décrit un outil logiciel permettant d'identifier des agglomérations dans des cadres à information limitée. Le cadre présenté dans la présente communication a été conçu pour fonctionner à l'aide de données ponctuelles. Notre outil, et toutes les données requises, sont fournis gratuitement et en format Open Source. *La présente communication décrit les avantages et les inconvénients de l'emploi de géographies de points dans l'analyse régionale, discute des difficultés pratiques et éthiques dans la distinction entre des agglomérations urbaines et des zones rurales, détaille la fonction de notre logiciel, et guide le lecteur intéressé vers le code source. La communication examine également les résultats de l'outil pour le Sri Lanka, et les compare avec les chiffres publiés par les Nations unies sur l'urbanisation. Nos résultats indiquent que la population urbaine du Sri Lanka est fortement sous-dénombrée dans les statistiques officielles.*

Jennifer Day, Faculty of Architecture, Building, and Planning, The University of Melbourne, Room 520, 757 Swanston St., Parkville, VIC 3010, Australia. Email: jday@unimelb.edu.au (to whom correspondence should be sent). Yiqun Chen, Faculty of Engineering, The University of Melbourne, Parkville, VIC 3010, Australia. Email: 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. Email: pellis@worldbank.org. Mark Roberts, The World Bank, 1818 H Street, NW Washington, DC 20433, USA. Email: mroberts1@worldbank.org

Una herramienta gratuita de código abierto para determinar las aglomeraciones urbanas utilizando los datos del punto

RESUMEN Este estudio describe a una herramienta de software que se utiliza para determinar las aglomeraciones urbanas en entornos donde hay poca información. La estructura que se describe en este estudio está diseñada para trabajar con los datos del punto. Nuestra herramienta y todos los datos requeridos se suministran de forma gratuita y en formato de código abierto. Este estudio describe las ventajas y desventajas de utilizar lugares geográficos basados en puntos en los análisis regionales, aborda los retos prácticos y éticos a la hora de diferenciar las regiones urbanas de las rurales, describe la función de nuestro software y dirige el interés del lector al código fuente. El estudio también analiza los resultados de la herramienta para Sri Lanka y los compara con las cifras sobre urbanización de las Naciones Unidas. Nuestros resultados indican que el recuento de la población urbana de Sri Lanka es significativamente menor en las estadísticas oficiales.

一种使用点数据辨识城市聚集的免费开源工具

摘要 本文介绍了一种用于在信息量少的情况下辨识城市聚集的软件工具。其中所概述的框架运用点数据。我们的工具和所有所需的数据均以开源形式免费提供。文中说明了在区域分析中使用基于点源的地理区域的优势及劣势，探讨了区分城市和农村区域所面临的实际和道德挑战，详细阐述了我们软件的功能，并为有兴趣的读者提供源代码。本文还分析了该工具生成的斯里兰卡输出结果，与公布的联合国城市化数据进行比照。我们的输出结果显示，官方统计数据远远低于斯里兰卡的城市人口。

Keywords: *Urbanization; urban extents; point data; metropolitan region*

JEL Classification: C81*;* J11*;* R58

1. Background and Motivation

This paper provides a framework and a functioning software script for identifying urban agglomerations in data-poor settings using free, open-source data. Our software tool, built in Cran R using data hosted in the public domain, is accessible to anyone with a sufficiently appointed computer and access to the internet. The software tool and framework outlined in this paper is designed to work using point data —that is, data attributed to a single geographic coordinate. We call this tool the Agglomeration Grid Algorithm.

Our overarching objective is to provide a method for bypassing established political designation of urban regions. These can confound cross-country comparisons and deny the existence of urban populations in areas officially classified as rural, among other issues. This paper has three main objectives. First, we argue for the importance of tools like this one and place it in a theoretical context. We then outline the decision structure of the algorithm we have created, providing the reader with enough information to understand how the software works. We then examine some outputs from the software and compare our outputs with published United Nations (UN) urbanization figures for Sri Lanka.

1.1. The Challenge of Characterizing Urban Regions

Urban regions are complex and diverse in many ways, including at their geographic edges. In some countries like China, urban agglomerations remain relatively discrete, surrounded by planned hinterlands into which the urban region expands over time (e.g. Day & Cervero, 2010). In places like the island of Java, Indonesia, however, continuous urbanization of varying intensity covers much of the island. The urbanized area around Jakarta, for instance, spills over the administrative Jakarta Special Capital City District boundary and into adjacent administrative units currently classified as rural.

McGee (1969) names this type of continuous urbanization, *desakota*, from the Indonesian words *kota*, meaning city, and *desa*, village. Rotgé (2001) identifies this type of urbanization in Yogyakarta on Java, and UNESCAP (2001) identifies it in Colombo, Sri Lanka. Jones (2001) and the World Bank (2012) suspect undercounts of the population that is economically and culturally tied to Jakarta.

Recognition of McGee's pattern of urbanization raises the question of whether classification of populations into urban and rural settlements is useful in analysis of development. Cohen (2004) cites technology diffusion and changes in livelihoods among other processes that blur the lines between urban and rural life. There are also arguments that formalizing urban/rural distinctions can entrench power structures and justify denial of services to some populations (e.g. Scott, 1998; Yiftachel, 1998).

On the other hand, there are practical issues for which development planners rely on data reflecting the distribution of urban and rural populations. Perhaps the most pressing issue has been reducing the disparity in health services, wealth and income, and infrastructure provision between urban and rural populations. Reducing disparities between urban and rural regions has been a sustained focus of development agencies (e.g. UNESCAP, 2001, p. 1; World Bank, 2009, 2012). As the debates suggest, and as Jonas and Ward (2007, p. 176) argue, cities are complex social, political, and economic constructions. We provide this tool recognizing that urban/rural distinctions are distortions of reality that can serve the interests of entrenched power (Scott, 1998). We think that tools like ours can assist in rebalancing that power.

Our tool, while providing a framework to identify urban agglomerations, does not address the underlying drivers of agglomeration as suggested by spatial economic theories such as urban economics and the New Economic Geography (NEG), for example, transport costs, industry returns to scale (Fujita & Thisse, 2003, p. 22), tension between centrifugal and centripetal forces (Krugman, 1996; Tabuchi, 1998), or knowledge transfer effects (Saxenian, 2000). Nor do we address the drivers of intra-metropolitan heterogeneity, for example, residential sorting according to public goods (Tiebout, 1956) or transportation and land rents (Alonso, 1975).

Rather, our purpose is to observe the overall urban footprint as evidence that these processes produce urban spaces in which people are concentrated, and to propose a framework to help analysts identify and differentiate them. The forces that compel cities into being, perpetuate their existence, differentiate them in size and industry, and cause wage and density differences within them are important, and we hope our tool will assist these types of queries. However, we start with a much more fundamental question: *Where are the cities?*

1.2. The Reality of Defining Urban Regions

Despite the widespread use of urban/rural distinctions in regional analysis, there is no widely adopted standard for defining metropolitan regions. The methods by which populations are classified as urban and rural vary widely among countries (Cohen, 2004). In Sri Lanka, for instance, urban and municipal councils range in population from over 600,000 in Colombo to as few as a few thousand. Required urban population counts vary, from 2,000 inhabitants in Angola and Ethiopia to 10,000 in Benin. Botswana classifies places as urban where there are 5,000 or more inhabitants and more than 75% of the economic activity is nonagricultural (Cohen, 2004). Urban designations within countries can also change over time. In China, for instance, official urban population more than doubled between 1982 and 1989—not because of a major population shift, but because the threshold at which a settlement was defined as urban changed in that period (UNESCAP, 2001, p. 5).

In research, the UN World Urbanization Prospects (WUP) dataset is perhaps the most widely cited measure of urban populations (World Urbanization Prospects, 2011). UN urbanization statistics, and research that uses them, rely on country-defined urban boundaries and population counts (e.g. Kalirajan & Otsuka, 2010). Analysts that do attempt to go beyond the UN data generally adopt *ad hoc*, rather than standardized, methods in their analyses (e.g. Balk et al., 2006; Cervero & Day, 2008; Day & Cervero, 2010; Day & Ellis, 2013; Day & Lewis, 2013). Other available data sources also do not entirely address the needs of regional analysis. Data provided by the Center for International Earth Science Information Network (CIESIN) provides information on urban extents in its Global Rural-Urban Mapping Project (GRUMP). This data source provides raster images of urban areas at a reasonably fine spatial scale, at a 2.5 arc-second grid (around 4.5 km on a side in Sri Lanka). However, distinguishing adjacent urban regions is impossible with this data.

This paper provides both a process for identifying urban regions and a software tool to perform the task. Our process is simpler than procedures used in high-income countries, which urban regions are often defined with more-sophisticated measures such as commute sheds (e.g. OECD, 2012, p. 23). Even in developed-country contexts, policy-makers continue to assert the need for more-refined spatial differentiation between urban and rural communities (e.g. Isserman, 2005). Recent work on Functional Urban Areas by the Organisation for Economic Co-operation and Development (OECD) states that a major constraint of its methodology for member countries is availability of commute data for smaller nations (OECD, 2012, p. 23).

1.3. Building on a New Practice for Defining Urban Regions

The process presented here draws from and extends Uchida and Nelson's (2009) Agglomeration Index. In their paper, the authors devise a standard process for identifying metropolitan agglomerations, with the intention of devising a globally consistent measure of settlement concentration for cross-country analysis. They then report their Agglomeration Index for around 200 countries.

Uchida and Nelson start by specifying three parameters that will be used to construct a metropolitan region: minimum population count, maximum travel time, and minimum population density. For them, a metropolitan region is an area that starts from the centre point of the city, is reachable in a certain travel time, and has a certain population density. They locate the centre point of cities with sufficient population using the GRUMP human settlements database. In this paper, we refer to this centre point as a *seed* or a *seed city* (we use the terms interchangeably), because it provides the base geography upon which the metropolitan region is constructed.

Once Uchida and Nelson have identified the centre point of a city, they use the second parameter—maximum travel time—to identify the furthest potential reaches of the metropolitan region. To do this, they develop a *cost surface* with a 1-km^2 grid. This cost surface reflects the time it takes to travel through that grid; allowing for the computation of a time-optimised path between places on the grid. They estimate on-road and off-road travel times based on road network, waterways, ground slope, and land cover.

Once they have computed the potential reaches of the region, Uchida and Nelson then apply the final test as to whether a grid cell is part of the metropolitan region: they allow grids within the maximum travel time threshold to be part of a metropolitan region if they meet a minimum population density. They use 1-km^2 spatial population data provided by two sources: the GRUMP gridded population data provided by CIESIN and Landscan, a dataset developed by Oak Ridge National Laboratory. They average the values for the two datasets because of their different methods used to project data from larger areas to smaller areas, reasoning that both datasets have advantages and limitations, and their average is likely better than either method alone. The outputs of their process are national urbanization rates, or urban population as a percent of national population.

Uchida and Nelson's algorithm is a very useful start, particularly for cross-country comparisons of urban populations, but it is not sufficient for spatial, metropolitan-scale analysis. Most importantly, their process is not a spatial one. The outcome of their process is the urban share of the national population, but not the spatial location of urban regions and their extents.

Because Uchida and Nelson's process is non-spatial, there are some significant problems to be addressed in applying it to spatial regional analysis. First of all, their process ignores issues of adjacency: if there are spatial gaps in the metropolitan region, their process does not identify them. Relatedly, their process provides no guidance on whether two large and adjacent cities should belong to the same metropolitan area, or alternatively, whether a large *kotadesa* region should be divided into two or more functioning metropolitan areas. Also, when two adjacent regions are identified, Uchida and Nelson's process does not provide guidance on how to allocate border regions between the two adjacent metropolitan areas. The latter two of these problems also occur in spatial datasets such as CIESIN's urban extents and nighttime lights data, which provide images of urbanized areas, but no boundaries between adjacent ones.

In addition, their introduction of travel time criteria to the process of identifying urban agglomerations is problematic because it does not take local traffic congestion or road conditions into account. Their travel times are estimated based on the physical configuration of the road, waterway, and rail networks as well as on walking speeds over the topography. For each of these modes, they estimate travel speeds, but these travel speeds are based on physical properties rather than non-physical attributes such as traffic and road conditions. Finally, although Uchida and Nelson provide a flexible framework within which to identify urban agglomerations, they stop short of providing a platform for the analysis. The next section outlines how we build upon their framework, addressing some of its shortcomings, to provide a new tool for identifying urban agglomerations.

1.4. Algorithm Overview

We start with Uchida and Nelson's notion that an urban agglomeration can be built by starting with a city with a certain population, and then adding places to the agglomeration according to whether they meet density and travel time criteria. We use the same population data that Uchida and Nelson used, CIESIN's open-access data on population distribution and urban settlements. From there, we depart from their work in a number of notable ways.

First, we provide open-source software tools that can be accessed by any user free of charge. Our software is programmed in Cran R, an open-source analysis platform. We make our source code freely available so that any user can modify it as their needs require. It is available at www.github.com/yiqunc/gridaa. Second, rather than defining a travel surface based on assumed travel speeds, we take advantage of the current availability of the free online data and tools provided by Mapquest and Google route information. This eliminates the need for proprietary datasets and assumed, imprecise travel time estimations.

Third, we provide a framework specifically targeted at distinguishing whether large, continuous, *desakota*-style conurbations should functionally be split into smaller, adjacent metropolitan agglomerations. Fourth, we provide guidance on dealing with adjacent metropolitan agglomerations and allocation of border regions that could belong to two or more urban agglomerations. This is a particularly important feature in places like Java, where *desakota*-type urban fabric makes adjacent agglomerations difficult to distinguish from one another. Finally, we allow the user to define key variables, such as the seed population, population density, and travel time thresholds.

Analysts looking for more-nuanced definitions of urban and rural can use this tool to identify a variety of urbanization forms along the *desakota* continuum. Where more-comprehensive data are available on economic characteristics, land use, and governance, the outputs of our tool could be integrated for spatial analysis. The remainder of this paper describes the features and underlying process of our algorithm and tool, and demonstrates its application to Sri Lanka.

2. Algorithm Design, Data, and Outputs

This section describes the elements of our Agglomeration Grid Algorithm, including the user-defined functions and the decision structures that a user interested in running the source code should understand. This section additionally details the data that the algorithm is capable of taking as inputs. Section 3 shows the output of the algorithm. All of the processes described here are coded using Cran R, a free and open-source analysis platform. Figure 1 shows the general logical flow of the algorithm, and in the remainder of Section 2, we describe the main features and methodological decisions in the tool.

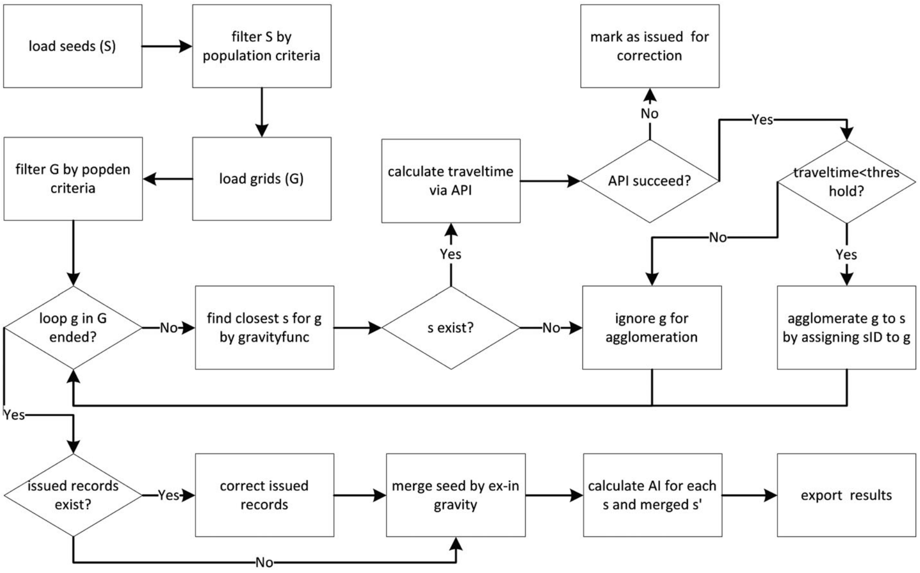
load seeds (S)
filter S by population criteria
mark as issued for correction
No
Yes
filter G by popden criteria
load grids (G)
calculate traveltime via API
API succeed?
traveltime<threshold?
No
Yes
Yes
loop g in G ended?
No
find closest s for g by gravityfunc
s exist?
No
ignore g for agglomeration
agglomerate g to s by assigning sID to g
Yes
issued records exist?
Yes
correct issued records
merge seed by ex-in gravity
calculate AI for each s and merged s'
export results
No

Figure 1. Agglomeration grid algorithm decision structure.

2.1. 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 tools for the construction of urban agglomerations. Our Agglomeration Grid Algorithm's user-defined parameters are summarized in Table 1, including the input

Table 1. Agglomeration grid algorithm user-defined parameters

Key parameters	Description	For Sri Lanka
gridFileName	Population grid file name (shp file POINT format). Generated by joining PCG with PDG, which are available in CIESIN GPW V3.	Use 2000 population data
seedFileName (optional)	Seed file name (shp file POINT format). Uses the Settlement Points V1 data available in CIESIN. If provided, the algorithm will use the *Settlement_Seeds* option; otherwise, it will use *Grid_Seeds* option.	Use 2000 settlement data
thPopDen	Threshold value indicating the minimum population density (people/km^2) at which a grid will be considered to be part of the metropolitan region.	150
thSeedMinPop	A threshold value indicating the minimum population at which an urban concentration will be considered as a seed.	50,000
thTvlTime	A threshold value indicating the maximum travel time (in minutes) between a grid and a seed for which the grid will be considered to agglomerate to the seed.	60
thSearchSeedRadius	A threshold value indicating the maximum Euclidian distance (in km) between a grid and a seed at which the seed will be considered as a candidate for the grid to attach to.	80
gridEdgeLength	The length of grid cell edge (in km). Used to determine how close a grid can be considered to contribute to a seed's internal gravity computation.	4.5

parameters we used to construct urban agglomerations for Sri Lanka that are presented in Section 3. The foregoing sections describe these input parameters in more detail.

2.2. Point Population Data

We start with two datasets: CIESIN's GRUMP grid of population count and population density, and CIESIN's list of urban settlements with associated populations. These datasets are available for most countries. CIESIN provides population data in roughly 4.5-km by 4.5-km^2 grid cells,[1] which they call the Population Count Grid (PCG) and Population Density Grid (PDG), with the data for the square area associated with a centroid point at the centre of each grid cell. The urban settlement data are the presumed centres of the urban administrative area (Balk et al., 2006), and are provided by CIESIN in point format in a dataset titled, Settlement Points. These settlement points act as the *seeds* for each urban agglomeration (seed designation is described in Section 2.3). The latest data for most CIESIN's datasets are around the year 2000. Figure 2 shows settlement points population count for these grid point centroids for Sri Lanka in 2000, with less-populated areas coloured in dark blue and more-populated areas in lighter shades of blue and in red. As we say above, this presentation of the data is useful in identifying the areas

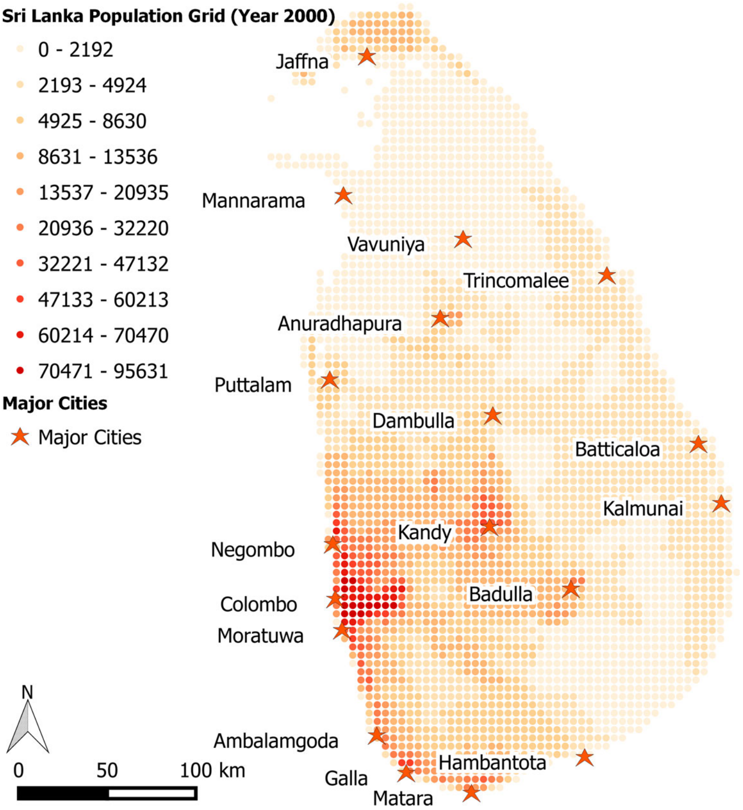

Figure 2. Sri Lanka population distribution.

that are urban, but not very useful in identifying the number and extents of individual agglomerations.

The data required to run our algorithm are minimal. These data include a list of potential seeds, the GRUMP point data described above, population within each seed area and grid cell centroid, and travel time. The source code we provide can be applied to any point-based dataset. The process by which we derive our seed cities is complex and warrants its own section. This appears below.

2.3. Seeds

As in Uchida and Nelson's process, our metropolitan areas start with a city centre point. We call this point a *seed*. Each agglomeration requires a seed, from which travel times are computed to neighbouring grid cell centroids and the parameters applied to test whether a that grid cell centroid shall be included in the agglomeration. Our software defines seeds according to a two-step process. It first identifies a set of *potential seeds*. Then, it applies a decision logic to determine whether *potential seeds* that are nearby each other should stand alone as the centre of a metropolitan area or be combined into a single seed, and thus a single metropolitan region. As we will demonstrate later in this paper, this is an important feature in countries like Sri Lanka, where large cities can be nearby each other.

Our software provides two options that an analyst can use to generate the potential seeds. The first option allows seeds to be identified directly from the GRUMP PCG dataset. Grid points meeting a certain population threshold are included in a list of *potential seeds*. Section 2.5 outlines the process by which some seeds are merged with other nearby seeds, in order to avoid unnecessary parsing of metropolitan agglomerations. We denote this as the *Grid_Seeds* option in Table 1.

The second mode our algorithm provides for generating seeds, we call the *Settlement_Seeds* option. In this mode, a list of potential seeds is comprised of a GRUMP settlement point from the Settlement Points dataset provided on the CIESIN website. Again, to be included in the list of potential seeds, the settlement point must contain a population above a certain threshold. We denote this option as '*Settlement_Seeds*' in Table 1 above.

In both the *Grid_Seeds* and *Settlement_Seeds* modes, the algorithm allows the user to define the relevant population threshold above which a point will be considered in the list of potential seeds. Uchida and Nelson use 20,000, 50,000, and 100,000 as thresholds, so we follow this process in producing our empirical outcomes below. We note that the two modes do to necessarily provide corresponding processes, as the *Settlement_Seeds* option is based on city population, and the *Grid_Seeds* option is based on the population of a much smaller area.

In our analysis, we define cities according to designations published by CIESIN, though a user could choose to define cities in other ways and input this data source into the software we provide. Table 2 lists cities in Sri Lanka as provided by CIESIN, and sorted by settlement size.

After a list of potential seed is designated, there is one more step in determining which of those seeds will act independently, and which will be combined with other seeds to form a single functional seed. We allow seeds to be merged into a single seed if they meet a merging criterion, which we describe in the next two sections. This process acknowledges that some metropolitan regions contain multiple cities with significant populations that are located in close proximity to each other. These

Table 2. Seed city by population, 2000

City	Population
Colombo	639,321
Dehiwala-Mount Lavinia	207,806
Moratuwa	174,786
Jaffna	128,001
Negombo	117,760
Kotte	115,038
Kandy	109,406
Kalmunai	94,364
Galla	90,173
Batticaloa	76,169
Katunayaka	70,019
Battaramulla	61,172
Dambulla	59,573
Anuradhapura	55,362
Kolonnawa	54,517
Maharagama	53,918
Kotikawatta	52,294
Daluguma	51,644
Trincomalee	47,977
Ratnapuraya	45,823
Matara	42,551
Mannarama	41,659
Badulla	40,486
Hendala	39,951
Puttalam	39,675
Kalutara	36,780
Keselwatta	36,703
Matale	35,990
Katankudi	34,666
Panadraya	33,311
Beruwala	32,677
Jaela	30,553
Peliyagoda	29,644
Welesara	29,016
Wattala-Mabola	28,328
Kurunegala	28,226
Homagama	28,042
Mulleriyawa	27,175
Nuwara Eliya	24,797
Gampola	24,119
Ragama	24,036
Chilaw	23,928
Kandana	23,494
Eravur	21,678
Weligama	21,635
Seethawakapura	21,207
Vavuniya	20,042

Source: CIESIN Settlement Points, 2000.

types of region configurations do not necessarily warrant parsing the region into a number of smaller, autonomous areas for analysis purposes. Section 2.6 describes this process.

2.4. Agglomeration Decision Structure

In defining metropolitan agglomerations, we needed to control for two types of problems that occur when there are multiple cities within close proximity to each other. First, some metropolitan regions contain multiple cities that meet a seed population threshold. The region around Colombo, Sri Lanka, for instance, has four cities with more than 100,000 people and eight cities with populations in excess of 50,000. 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 four or eight (depending on parameter specification) metropolitan regions by the algorithm. Not every city/seed is deserving of being identified as the core of a metropolitan region, as many metropolitan regions around the world contain a primate city and a number of smaller cities. We call this problem *underagglomeration*.

On the other hand, failing to recognize some non-primate cities as being the core of an autonomous metropolitan area is also a problem for this process. We do not wish to mistake two adjacent metropolitan regions as a single region. We call this problem *overagglomeration*.

To balance between overagglomeration and underagglomeration, we develop an algorithmic decision structure that tests whether agglomerations should be sub-divided. Uchida and Nelson's framework works by allowing each seed to 'claim' those grid points that meet the decision travel time and population density criteria. Their process starts at the seed and agglomerates grid points to the seed based on whether they meet population density and travel time conditions.

Our process achieves the same effective result, but works in the reverse of Uchida and Nelson. The algorithm starts with an arbitrary grid point g_i, which meets a user-specified minimum population density criterion. Let's say there are 1,000 data points in our grid, and 400 of them meet the population density threshold specified by the user. For each of those 400 g_i's, the algorithm then searches the candidate seeds to which that g_i could be assigned. To increase computational efficiency and to conform to practical reality, the user may limit the distance within which each grid point g_i can search for seeds. This distance is a user-specified distance threshold, which we call *thSearchSeedRadius* in Table 1.

Once all candidate seeds are found for each grid point, we need a decision structure to decide which agglomeration a given grid point will be assigned. This software assumes that each grid point can be assigned to only one agglomeration. For the demonstrations in this paper, we use a gravity index for this decision. Gravity indices, which use population and separation to estimate the attractive capacity of one place to another and the attenuation of that attraction due to distance, are common in regional analysis, for example, for analysis of economic growth spillover effects (Day & Lewis, 2013), to estimate regional job accessibility (Cervero et al., 1999), and to estimate urbanization effects on factor productivity (Beeson, 1987). Gravity models, which incorporate gravity measures in econometric estimations, are also common; for instance, in analysis of trade flows (Burger et al., 2009) and migration (Crozet, 2004).

Here, we estimate gravity with the following formula:

$$\text{GRAVITY}_i = \frac{\text{POPULATION}_j}{(\text{DISTANCE}_{\text{ij}})^2}, \qquad (1)$$

where GRAVITY$_j$ is the gravity measure representing the attractiveness of grid *i* to seed city *j*, POPULATION$_j$ is the population of seed city *j*, and DISTANCE$_{ij}$ is a measure of separation between grid *i* and seed city *j*. For this analysis, the gravity index measure is computed between each grid point g_i and all of its candidate seeds. Each grid point is assigned to the seed city that yields the highest gravity measure. We measure DISTANCE using the method described in Section 2.5.

In addition to the gravity measure given in Equation (1), our software tool has three other equations that a user can elect to use instead. These include attenuation measures based on distance only, population only, and population divided by distance rather than distance squared, as above. Users can also write their own equations into the software. We use Equation (1) in this paper because it produces the most interpretable results, and also because the process is theoretically defensible, since it considers both distance and size in a measure of attractive force. The squared term places more weight on the distance term.

These *intermediate agglomerations* are not the final step in the formation of agglomerations. If we left the formation of agglomerations here, we run the risk of underagglomerating (we describe the final step in the next section). This intermediate process helps us to avoid a major pitfall of Uchida and Nelson's method: their process can sometimes result in contested grid data points claimed by more than one seed, particularly in the border regions of two adjacent metropolitan areas. Our process assigns each grid point to only one seed, and so we avoid the problem of contested grid points.

2.5. Delineating Agglomerations

The algorithm logic described in the preceding section assures that there is no spatial overlap between agglomerations; however, it does not preclude underagglomeration. This tool assists the user in distinguishing whether a large urban conurbation should really be logically split into two or more adjacent but functionally separate agglomerations. Whenever an agglomeration contains two or more seeds, we apply a decision process to determine whether they should be allowed to maintain separated status, or whether their seeds should be 'merged' (we define a merger of two seed points below).

We reason that a region should remain functionally separate from a region it overlaps, if its internal cohesiveness is large enough to overcome the attraction of the larger region. Figure 3 illustrates this concept. These two hypothetical agglomerations are developed using Seed City A with 200,000 residents and Seed City B with 100,000 residents, and with density thresholds of 150 people/km^2 and a travel time threshold of 60 min. Agglomeration A, centred around Seed City A, is the larger (in population size) of the two agglomerations, and Agglomeration B is the smaller and centred around Seed City B.

Because Seed City A is larger and the two regions are adjoining, we wish to test whether Seed City B can be considered a functionally separate unit. To do this, we measure the *internal cohesion* of Agglomeration B and compare it to the *external cohesion* between Seed City A and Seed City B. If the internal cohesion in Agglomeration B is larger than the external cohesion between Seed Cities A and B, then Agglomeration B remains a separate agglomeration. If not, then Seed Cities A and B are 'merged' (the process of merging seeds is described in Section 2.5). To measure internal and external cohesion, we also use a gravity index. Figure 3

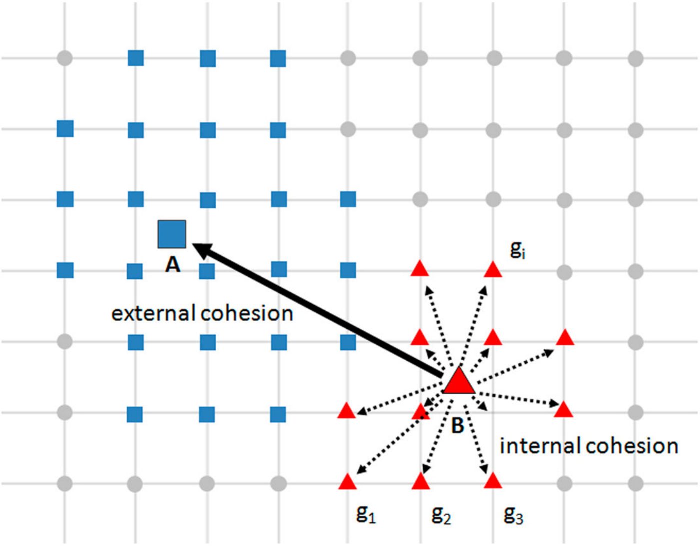

Figure 3. Internal versus external cohesion in the agglomeration grid algorithm.

illustrates the comparison of internal cohesion in Region B with the external cohesion between Region A (in blue/squares) and Region B (in red/triangles). We estimate the external cohesion between Seed Cities A and B, with the population of Seed City A being larger than the population of Seed City B, as:

$$\text{EXTERNAL\_COHESION}_{\text{A,B}} = \frac{\text{POPULATION}_{\text{A}}}{(\text{DISTANCE}_{\text{A,B}})^2}. \qquad (2)$$

Next, we estimate the internal cohesion of Agglomeration B. The internal Cohesion of Agglomeration B is the summed gravity measure of all points contained within Agglomeration B, to Seed City B:

$$\text{INTERNAL\_COHESION}_{\text{AggB}} = \sum_{k=1}^{n} \frac{\text{POPULATION}_{k}}{(\text{DISTANCE}_{k,\text{B}})^2}, \qquad (3)$$

where there are n grid points internal to Agglomeration B. Because this measure of internal cohesion is the summation of n gravity measures, and because seeds do not match up with the 4.5-km grid pattern, it is sometimes the case with this type dataset that grid points can be very close to seed points. This can inflate some gravity measures to very large magnitudes. In this case, a computed average would be inflated. To mitigate this problem, we drop all points k from the internal cohesion metric, that are within 4.5 km (one grid length) of a seed. Finally, our algorithm employs the following decision process to determine whether Agglomeration B should remain separate from Agglomeration A:

(i) if $\text{INTERNAL\_COHESION}_{\text{AggB}} > \text{EXTERNAL\_CHOHESION}_{\text{A,B}}$, Agglomeration B remains a separate agglomeration

(ii) if $\mathrm{INTERNAL\_COHESION}_{\mathrm{AggB}} <= \mathrm{EXTERNAL\_CHOHESION}_{\mathrm{A,B}}$, Seed City B ‘merges’ with Seed City A

If the conditions in (i) are met, Agglomeration B is allowed to remain as a separate agglomeration. The two terms, internal and external cohesion, are compatible even though one is a summation and one is not, because the population and distances are smaller for internal versus external cohesion. In this case, there can still be overlap between Agglomeration A and Agglomeration B. Since it is convenient in regional analysis for agglomerations to remain mutually exclusive, our algorithm contains a decision process that allocates each grid point to one of the seed cities. In Section 3.3, we demonstrate the effect that this decision structure has on agglomeration configurations.

Section 2.6 outlines how contested grid points are allocated to the appropriate seed. If the conditions in (ii) are met, we conclude that Seed City B is functionally a part of Agglomeration A. Seed Cities A and B are ‘merged’, and the Agglomeration Grid Algorithm is run again on the single new merged seed. The next section discusses how two seeds are ‘merged’.

2.6. Seed Merging

At the end of Section 2.3 and again at the end of Section 2.4, we introduce the idea that two seed cities can ‘merge’ if they meet certain merging criteria. We allow merger of seeds if the spatial structure of the data give us good reason to suspect that two seed points (whether generated with the grid or settlement process) are functionally part of the same metropolitan agglomeration. We use the term ‘merge’ in quotation marks because it is an antithetical concept in Euclidean point geometry. It is common to think of two adjacent polygons being merged into a single polygon because they share a common side. For points, however, there is no such thing as adjacency. It is an axiom of Euclidean geometry that for any two points, there can be placed a third point between them. The idea of ‘merging’, then, is not intuitive when the seeds are represented by points, as they are in the data structures presented here. The decision structure for merging seeds is described in the remainder of this section.

We begin with a set of all eligible seeds (those with populations greater than the user-defined threshold, *thSeedMinPop* from Table 1). The algorithm outlined in Section 2.4 is run on these candidate seeds. If condition (ii) from Section 2.4 above is met, then the two seeds are merged and the algorithm is run again, this time with one fewer seed than previously. The process starts with the two highest seeds in proximity to each other, and continues iteratively until all seeds are test for merger and merger is either allowed or rejected.

We mean something very specific when we say that seeds A and B are ‘merged’ in this software—and more specifically, when Seed B is merged to Seed A. When B merges to A, all grid points (g_i) previously allocated to B become now allocated to A instead, and B is dropped off the list of eligible seeds. We do not mean a merger in a spatial sense, where attributes such as population of the smaller seed are combined with those of the larger seed, and the new merged entity is considered to be spatially centred at the location of the larger seed.

Once the two seeds’ grid points (g_i) are combined, the software then runs the Agglomeration Grid Algorithm again, this time with one fewer seed than in the previous iteration. This process is repeated for all adjacent regions.

2.7. Allocation of Contested Border Points

In Section 2.4, we noted that when the conditions in (i) are met and Agglomeration B is allowed to remain as a separate agglomeration, the Agglomeration Grid Algorithm can sometimes result in contested grid data points claimed by more than one seed. This occurs in the border regions of two adjacent metropolitan areas. Since it is convenient in regional analysis for agglomerations to remain mutually exclusive, our algorithm contains a decision process that allocates each grid point to one of the seed cities.

In a case where two or more seeds claim a point, a decision algorithm is applied to assign the point to only one seed. This decision algorithm is based on a gravity measure in Equation (1). For each grid point that the algorithm assigns to more than one seed, a gravity measure is computed to estimate the attraction of that grid point to the seeds in question. For instance, if grid point i is allocated in the initial run to both Agglomeration A and Agglomeration B, then $\text{Gravity}_{i,\text{A}}$ and $\text{Gravity}_{i,\text{B}}$ are both computed. The grid point is then assigned to the region whose seed has the stronger attractiveness to it, that is, the grid point is assigned to the seed which results in a larger gravity measure. The point 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 3.

2.8. Travel Times

Travel times are computed using two free and navigation Application Programming Interfaces (APIs) provided as web services by MapQuest and Google. These APIs generally provide navigation and travel time data for most cities in most countries. This approach is free, fast, generally up-to-date, and considers traffic and road construction in many places. In this research, we develop a Cran R script to automatically interface with the APIs, so no additional steps or expertise are required by the user to obtain this travel data.

The MapQuest navigation API is called first because it allows unlimited requests. Google navigation API is invoked when the MapQuest cannot provide a travel time estimate. Google is called second because its API has query limit of 2,500 requests in 24 h. In the small number of instances (less than 5% in Sri Lanka) where neither Mapquest nor Google can generate travel times, the mean travel speed for all computed centroid-to-centroid segments is applied to the missing segment to arrive at an estimated travel time.

2.9. Algorithm Outputs

The outputs of the software tool include an ESRI Shapefile and a spreadsheet in comma-separated values (CSV) format. In the Shapefile, and also in the CSV file, each geographic point is associated with a numeric code identifying the metropolitan agglomeration to which it belongs. There is also a second numeric code, a dummy variable, which designates whether the point is a seed. We also include a unique identification number, population counts, and population densities. From these files, a user can generate maps and descriptive statistics outside of the software.

3. Algorithm Performance—Sri Lanka

This section describes the application of our Agglomeration Grid Algorithm to Sri Lanka. We compare the outputs of the algorithm under different parameter specifications for the year 2000, the latest year available from CIESIN. Subsequently in this paper, we use the following parameter combination shorthand to refer to parameter combinations: (150; 50,000; 60) refers to a minimum population density of 150 persons/km^2, 50,000 minimum seed population, and 60 min maximum travel time.

3.1. Algorithm Estimations of National Urbanization Rates

Table 3 shows the computed urbanization rates for Sri Lanka for 2000 under various parameter configurations, including both the *Settlement_Seeds* and the *Grid_Seeds* options for defining seeds. In crafting the parameters to be included in Table 3, we began with Uchida and Nelson's (2008) working specifications of 150 people/km^2, 50,000 urban inhabitants, and 60 min of travel time, or (150; 50,000; 60) in our shorthand. We then varied different parameters to observe how the algorithm outputs vary. We included different population thresholds and densities in recognition of the different urban forms and designations we discuss

Table 3. Urbanization rates under the *Settlement_Seeds* and *Grid_Seeds* options

Agglomeration algorithm parameters				2000 Agglomeration algorithm output			
Population density (people/km^2; minimum)	Seed population (minimum)	Travel Time (minutes; maximum)	2000 National population	Urban population	% Urban	Urban population	% Urban
				Settlement_Seeds Option		Grid_Seeds Option	
150	20,000	60	18,923,627	12,970,803	68.54	12,086,492	63.87
150	20,000	90	18,923,627	15,005,187	79.29	13,699,157	72.39
150	50,000	60	18,923,627	10,853,458	57.35	8,477,277	44.8
150	50,000	90	18,923,627	13,357,324	70.59	10,721,246	56.66
150	100,000	60	18,923,627	8,642,179	45.67	NA	NA
150	100,000	90	18,923,627	10,943,711	57.83	NA	NA
300	20,000	60	18,923,627	11,581,655	61.20	11,221,489	59.3
300	20,000	90	18,923,627	12,952,100	68.44	12,312,452	65.06
300	50,000	60	18,923,627	9,778,409	51.67	8,212,628	43.4
300	50,000	90	18,923,627	11,717,834	61.92	10,110,847	53.43
300	100,000	60	18,923,627	8,349,080	44.12	NA	NA
300	100,000	90	18,923,627	10,177,064	53.78	NA	NA
500	20,000	60	18,923,627	9,205,773	48.65	8,941,029	47.25
500	20,000	90	18,923,627	9,671,494	51.11	9,115,674	48.17
500	50,000	60	18,923,627	8,549,191	45.18	6,893,504	36.43
500	50,000	90	18,923,627	9,237,523	48.81	7,978,029	42.16
500	100,000	60	18,923,627	7,514,407	39.71	NA	NA
500	100,000	90	18,923,627	8,307,132	43.90	NA	NA
Summary statistics							
Mean				10,489,685	55.43	9,980,819	52.74
Minimum				7,514,407	39.71	6,893,504	36.43
Maximum				15,005,187	79.29	13,699,157	72.39

in the Section 1.1. Urbanization in Sri Lanka tends to follow a low-grade *desakota* pattern, so smaller thresholds were warranted. We include a larger travel time threshold in recognition of the road conditions in the country.

Depending on the parameters applied to our algorithm, year-2000 national urbanization rates vary between 40% and 79% when settlement points are used as seeds (8.5–15 million people), and from 36% to 72% when grid points are used as seeds (6.9–13.7 million people). This is indeed a large range. We provide these results to showcase the flexibility of the tools, and also to remind the reader that local knowledge can assist greatly with the selection of appropriate parameters in using this tool. Parameters should be adjusted according to local conditions. In a World Bank (2012) report on urbanization and growth in Indonesia, for instance, the authors used a preliminary process to the one described in this paper and arrive at the density thresholds for Java and other Indonesian islands using the 85th percentile of urban densities.

We note that although we use the same parameter configurations for both the *Settlement_Seeds* and the *Grid_Seeds* options, the results are not directly comparable because of the different seed configurations. Settlement points represent the entire urban agglomeration as defined by the country, while the grid seeds are based on the population of a single grid within a metropolitan area. We also note that no grid cells in Sri Lanka had populations greater than 100,000 persons, so no data points appear for rows using this configuration. The algorithm estimates higher urbanization rates when settlement points are used as seeds.

In contrast to the above ranges, the UN-estimated urbanization rate of Sri Lanka in 2001 is 14.58%. These are based on Sri Lankan Census data from 2001. One reason for this mismatch could be that the 2001 Census does not provide enumerations for some Divisional Secretariat-Divisions in the Northern and Eastern Provinces, due to the civil conflict in Sri Lanka at that time. However, the non-enumerated cities are far smaller than Colombo and could not account for a drop in the urbanization rate from even 36% to 14.58%. Even given a possible under-count of urban populations, then, this comparison makes clear that the algorithm estimates much higher urban populations percentages than do official statistics. The outputs of the Agglomeration Grid Algorithm, then, are aligned with estimates by other analysts who suggest that the urban populations in South Asia are under-counted (Cohen, 2004).

3.2. Number and Size of Agglomerations

The number and size of agglomerations varies depending on the parameter configurations. Figures 4 and 5 show the proportion of population and land area, respectively, contained in the overall rural area and in the metropolitan regions formed under six parameter combinations and using the *Settlement_Seeds* option. We note that all of the slices of the pies are not labelled. These figures show the general dynamics of the algorithm: larger seed population requirements result in fewer and larger metropolitan regions. This indicates that the *internal cohesiveness* feature of the algorithm is functioning. Smaller places can act as seeds, and the relative cohesiveness of their surrounds outweighs the attraction to larger places. Figure 4 indicates that increasing travel time thresholds will increase the estimated overall urbanized population. Under all configurations, Colombo stands out as a significant primate city, with Kandy a

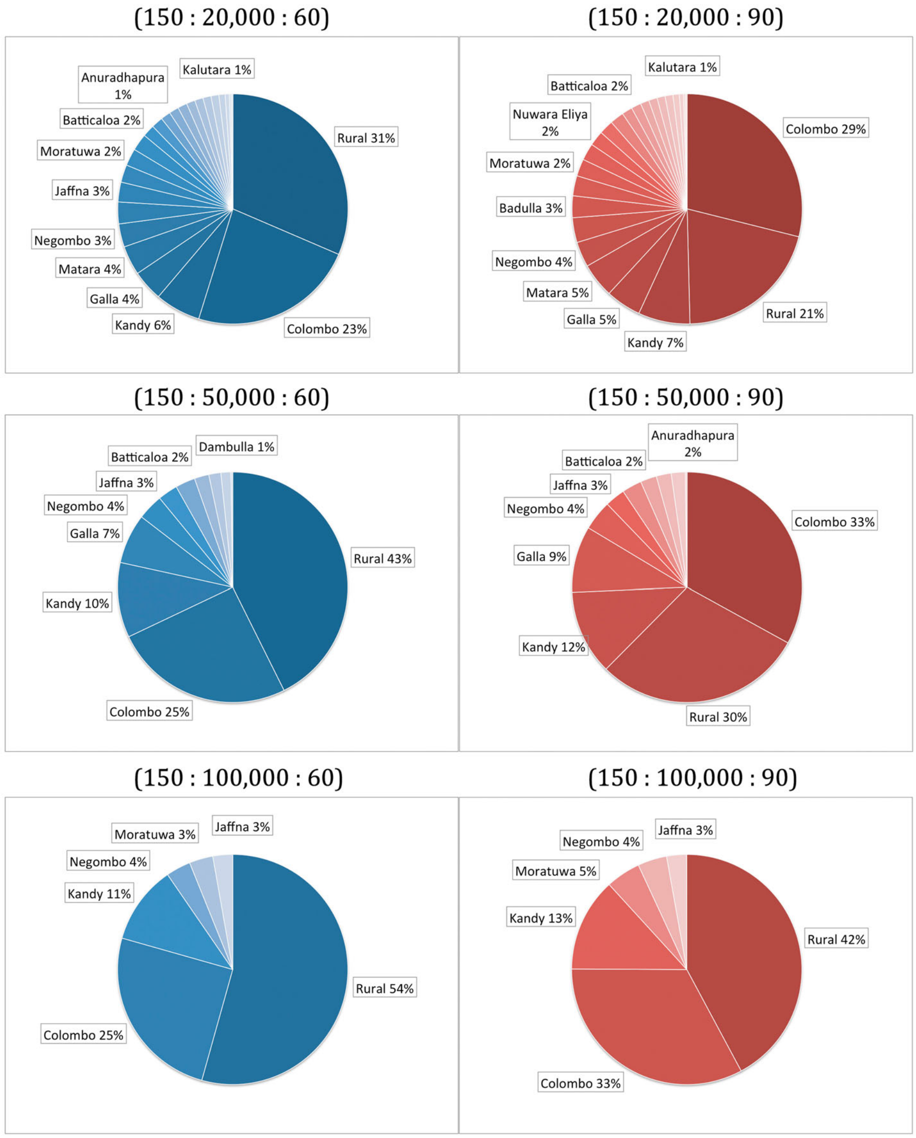

Figure 4. Percent of population living in metropolitan agglomerations and rural areas, 2000, selected parameter configuration.

distant second place. Table 4 shows the population and land area in urban regions under the parameter combination (150; 50,000; 60).

3.3. Agglomeration Configurations

Figure 6 shows the spatial configuration of the algorithm-computed agglomerations under four parameter configurations: (150; 50,000; 60), (150; 50,000; 90), (150; 100,000; 60), and (150; 100,000; 90). As a naming convention, we use the name of the largest seed city in the agglomeration.

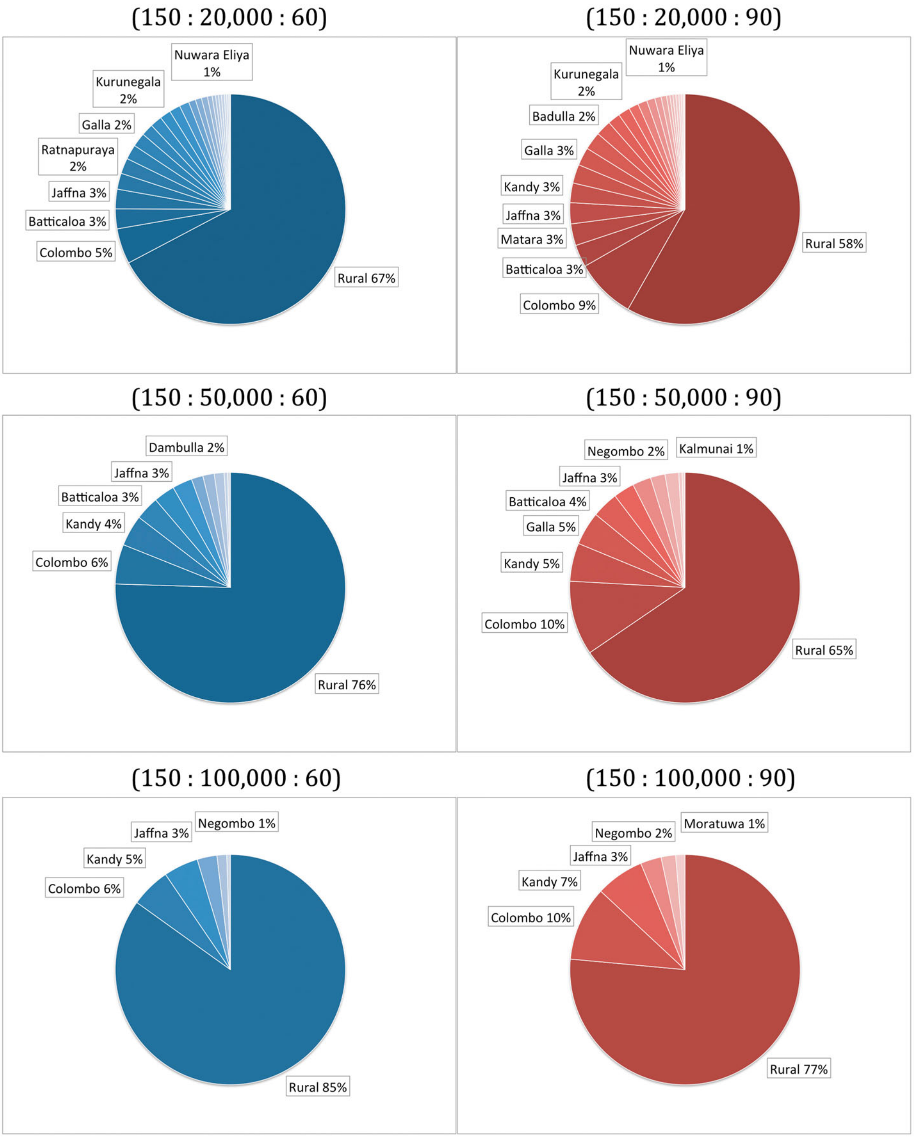

Figure 5. Percent of land area contained in metropolitan agglomerations and rural areas, 2000, selected parameter configurations.

All four of these maps indicate that a large portion of the south-west coast of the country is characterized by adjacent and large metropolitan regions. This is consistent with McGee's *desakota*-type urbanization. As expected, the estimated urbanized areas increase with travel time and decrease with larger seed requirements.

In Sections 2.4 and 2.5, we introduce our algorithm decision structure that allows seeds to be merged to prevent underagglomeration. Here, we demonstrate the results of this decision structure, visually. Figure 7 shows detail of the

Table 4. Sri Lanka metropolitan agglomerations in the (150; 50,000; 60) configuration

Agglomeration	Population	Population (%)	Area (%)
Colombo	4,787,286	25.30	5.57
Kandy	1,986,480	10.50	4.36
Galla	1,324,445	7.00	2.96
Negombo	683,838	3.61	1.39
Jaffna	526,637	2.78	2.79
Moratuwa	526,052	2.78	0.39
Batticaloa	384,883	2.03	3.35
Anuradhapura	320,994	1.70	1.63
Dambulla	264,767	1.40	1.60
Kalmunai	48,077	0.25	0.47
Urban total	10,853,458	57.35	24.51
Rural	8,070,169	42.65	75.49
Sri Lanka total	18,923,627	100.00	100.00

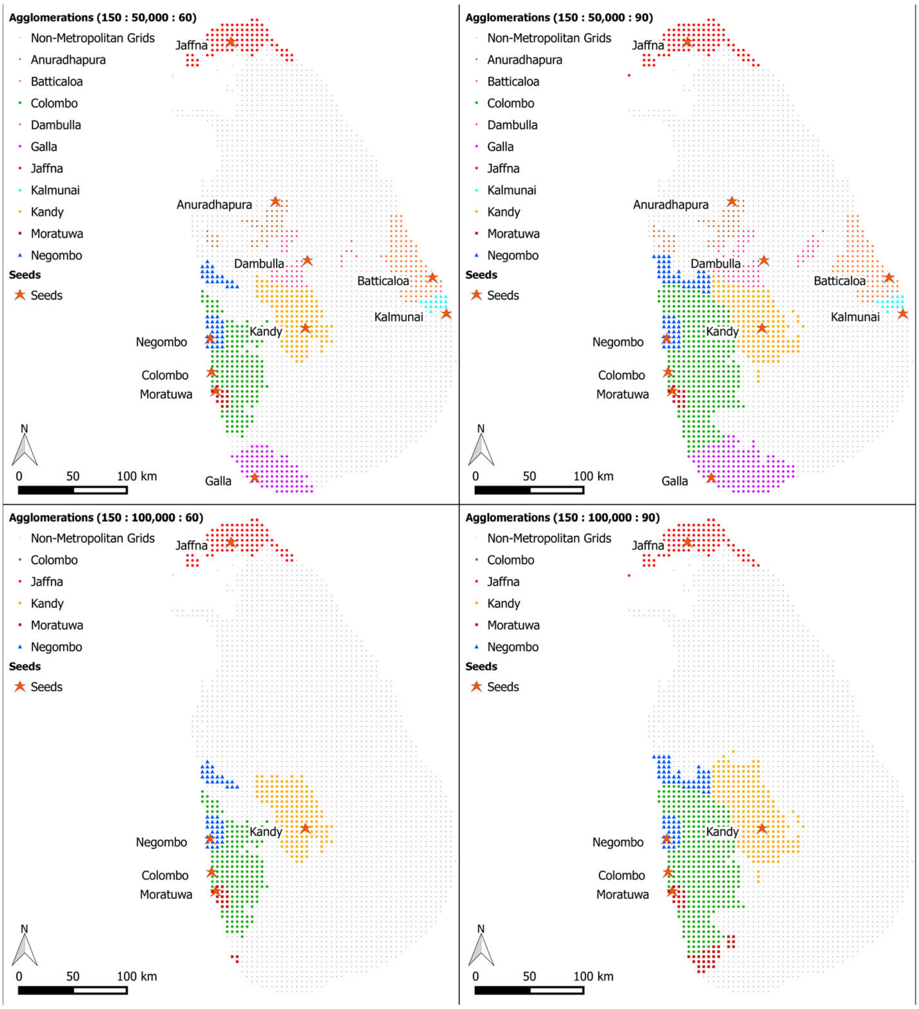

Figure 6. Agglomerations under four parameter specifications, settlement points as seeds.

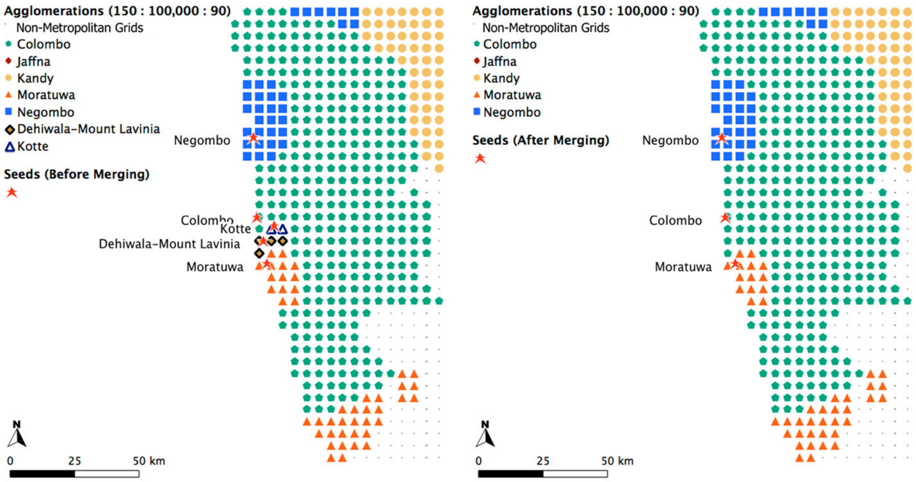

Figure 7. Colombo detail before and after seed merging, settlement points as seeds.

Colombo metropolitan region under a single parameter configuration: (150; 100,000; 90). The first map depicts the original agglomerations around the south-west coast, before the seed merger process is applied. The second map depicts the further agglomerations after merging seeds using the external–internal gravity decision structure outlined in Section 2.4. We note that in the original configuration, Kotte and Dehiwala-Mount Lavinia are the centres of independent agglomerations. In the second map, they are absorbed by Colombo.

This algorithm can result in some peculiar spatial configurations of agglomerations. In some specifications, some metropolitan regions are not spatially contiguous, or one region may be contained entirely within another. The Moratuwa and Negambo metropolitan regions are not spatially contiguous under two of the four parameter specifications mapped in Figure 6, (150; 100,000; 60) and (150; 100,000; 90). Also, both Figures 6 and 7 indicate that the regions of Moratuwa and Negambo are completely surrounded by grid points allocated to the Colombo region.

The reason for these counter-intuitive spatial patterns lies in the use of the gravity index described above. This measure takes into account travel time and population attraction of a seed region to a particular grid point. Because of the large population mass of Colombo, grid points must be sufficiently close to a smaller seed like Moratuwa, or else sufficiently far from Colombo, in order to be allocated to smaller seeds.

These unusual configurations illustrate the difficulty of a universal process without local knowledge, and make one wonder how many such discrepancies exist in Uchida and Nelson's process. As one can see from Table 1, Moratuwa is larger in population than Negombo. However, locally, Moratuwa is considered to be a suburb of Colombo, while Negombo is considered to be a city in its own right. Local analysts could exclude Moratuwa from the list of seeds; we could not make this decision based on the CIESIN settlement points data alone.

3.4. Comparisons with UN Agglomeration Definitions

Using the WUP urban-size designations (World Urbanization Prospects, 2011), we compare urbanization rates from our algorithm with WUP data for the year 2000. The WUP data only catalogues agglomerations with populations greater than 500,000, so we cannot compare the data for agglomerations smaller than a half-million people. Still, the comparison for larger classes of cities provides context for interpreting the AI estimates.

Table 5 gives the number of agglomerations, urban population, and percent of urban population held in each of five city size classifications and overall. Urbanization estimates are given using the Agglomeration Grid Algorithm and the WUP data. Table 5 shows results based on the (150; 50,000; 60) configuration.

Perhaps the most striking contrast between the WUP data and the algorithm-estimated urbanization data is in the number of people classified as urban dwellers. WUP estimates that 2.9 million people reside in urban regions; the algorithm estimates 10.8 and 7.3 million urban residents in Table 5. Another notable difference occurs in the high-population agglomerations. The WUP data do not classify any urban agglomerations as having more than one million persons. The algorithm, in contrast, attributes a high proportion of the urban population to a few large urban centres with between one and five million people. This difference appears to be at least partially the result of more people being located in larger metropolitan

Table 5. Sri Lanka comparison of computed urbanization characteristics, 2000, with WUP statistics, 2000 (150; 50,000; 60) configuration

City size	Parameter	WUP	Agglomeration grid algorithm	Difference
10 million or more	Number of agglomerations	0	0	0
	Percentage of urban population	0	0	0
	Population	0	0	0
5–10 million	Number of agglomerations	0	0	0
	Percentage of urban population	0	0	0
	Population	0	0	0
1–5 million	Number of agglomerations	0	3	3
	Percentage of urban population	0	74.6	74.6
	Population	0	8,098,211	8,098,211
500,000 to 1 million	Number of agglomerations	1	3	2
	Percentage of urban population	22.0	16.0	−6.0
	Population	644,000	1,736,527	1,092,527
Less than 500,000	Number of agglomerations	–	4	–
	Percentage of urban population	78.0	9.4	−72.7
	Population	2,301,000	1,018,721	−1,739,971
Total	Number of agglomerations	–	10	–
	Percentage of urban population	100.0	100.0	0.0
	Total urban population	2,945,000	10,853,458	7,908,458

Notes: (–) denotes data not available or not computable.

areas with the algorithm. WUP estimates that 78% of Sri Lanka's urban dwellers live in cities with fewer than 500,000 people. These two algorithm configurations attribute a much smaller population to small regions: 9.4% of urban population in Table 5.

We realize that these differences raise questions about the viability of our algorithm. However, much more importantly, they also raise questions about the viability of the most widely cited data on urbanization. This is a conversation that the community of urban scholars needs to have.

4. Conclusions

This paper describes a framework for identifying metropolitan agglomerations, and for distinguishing contiguous agglomerations into functional ones. The tool, extending the framework provided by Uchida and Nelson (2009), could be useful for urban and regional analysis in countries where sophisticated data is not available to define metropolitan extents, and in cross-country analysis.

Given the different and widely varying urbanization estimates and configurations that different parameter specifications provide, we recognize a need to provide an analytical framework for choosing parameter configurations appropriate to the country setting. At present, our tool does not provide an analyst with a way to select the best set of parameters for the analysis setting. The problem in identifying an optimum set of parameters is that there is no benchmark dataset of 'true' metropolitan designations against which we can test our outcomes.

Despite this limitation, our tool moves toward mitigation of another significant, if largely hidden, problem of false precision in the state of current analysis that seeks to identify urban regions. The current practice of relying on government-defined boundaries presumes that their processes are sound. Our tool provides the process for critique and adjustment. We hope that a community of analysts and policymakers will test the framework's applicability in applied settings and built on our process toward producing a more-universally applicable system for identifying reliable metropolitan extents.

Acknowledgements

The development of this tool was motivated by recent and ongoing work by the World Bank for the South Asia region, where new processes were required to identify urban regions. This work was not funded by the World Bank. Thanks to Pietro Calogero, Sophie Sturup, Ria Hutabarat, and Stefan Al for their comments and suggestions on the writing.

Disclosure statement

No potential conflict of interest was reported by the authors.

Note

[1] More precisely, CIESIN provides 2.5 arc-minute grid cells, and in Sri Lanka, a single grid size roughly equals 4.5 km by 4.5 km

References

Alonso, W. (1975) Location theory, in: J. Friedman & W. Alonso (eds) *Regional Policy: Readings in Theory and Applications*, pp. 35–63, Cambridge, MIT Press.

Balk, D. L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S. I. & Nelson, A. (2006) Determining global population distribution: methods, *Applications and Data, Advances in Parasitology*, 62, 120–151.

Beeson, P. (1987) Total factor productivity growth and agglomeration economies in manufacturing, 1959–73, *Journal of Regional Science*, 27(2), 183–199.

Burger, M., Van Oort, F. & Linders, G.-J. (2009) On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation, *Spatial Economic Analysis*, 4(2), 167–190.

Cervero, R. & Day, J. (2008) Suburbanization and transit-oriented development in China, *Transport Policy*, 15(5), 315–323.

Cervero, R., Rood, T. & Appleyard, B. (1999) Tracking accessibility: employment and housing opportunities in the San Francisco Bay Area, *Environment and Planning A*, 31(7), 1259–1278.

Cohen, B. (2004) Urban growth in developing countries: a review of current trends and a caution regarding existing forecasts, *World Development*, 32(1), 23–51.

Crozet, M. (2004) Do migrants follow market potentials? An estimation of a new economic geography model, *Journal of Economic Geography*, 4(4), 439–458.

Day, J. & Cervero, R. (2010) Effects of residential relocation on household and commuting expenditures in Shanghai, China, *International Journal of Urban and Regional Research*, 34(4), 762–788.

Day, J. & Ellis, P. (2013) Urbanization for everyone: the benefits of urbanization in Indonesia's rural regions, *Journal of Urban Planning and Development*, 140(3), 04014006–1–04014006–9.

Day, J. & Lewis, B. (2013) Beyond univariate measurement of spatial autocorrelation: disaggregated spillover effects for Indonesia, *Annals of GIS*, 19(3), 169–185.

Fujita, M. & Thisse, J. F. (2003) Does geographical agglomeration foster economic growth? And who gains and loses from it?, *Japanese Economic Review*, 54(2), 121–145.

Isserman, A. M. (2005) In the national interest: defining rural and urban correctly in research and public policy, *International Regional Science Review*, 28(4), 465–499.

Jonas, A. E. & Ward, K. (2007) Introduction to a debate on city-regions: new geographies of governance, democracy and social reproduction, *International Journal of Urban and Regional Research*, 31(1), 169–178.

Jones, G. W. (2001) Studying extended metropolitan regions in South-East Asia. Paper presented at the International Union for the Scientific Study of Population, 2001 Meeting, Salvador, Brazil.

Kalirajan, K. & Otsuka, K. (2010) Decentralization in India: outcomes and opportunities: Australian National University (ANU), Australia South Asia Research Centre (ASARC).

Krugman, P. (1996) Urban concentration: the role of increasing returns and transport costs, *International Regional Science Review*, 19(1–2), 5–30.

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*, pp. 93–108, Honolulu, University of Hawaii Press.

OECD (2012) Redefining urban: a new way to measure metropolitan areas; functional urban areas in OECD Countries, Organisation for Economic Cooperation and Development.

Rotgé, V. (2001) *Rural-urban Integration in Java: Consequences for Regional Development and Employment*, Aldershot, Ashgate Publishing.

Saxenian, A. (2000) Regional networks and innovation in Silicon valley and route 128, in: Z. J. Acs (ed) *Regional Innovation, Knowledge and Global Change*, pp. 123–138, London and New York, Pinter.

Scott, J. C. (1998) *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*, New Haven, Yale University Press.

Tabuchi, T. (1998) Urban agglomeration and dispersion: a synthesis of Alonso and Krugman, *Journal of Urban Economics*, 44(3), 333–351.

Tiebout, C. M. (1956) A pure theory of local expenditures, *The Journal of Political Economy*, 64(5), 416–424.

Uchida, H. & Nelson, A. (2009) Agglomeration index: towards a new measure of urban concentration, world development report 2009, The World Bank, Washington, D.C.

UNESCAP (2001) Reducing disparities: balanced development of urban and rural areas and regions within the countries of Asia and the Pacific, United Nations Economic and Social Commission for Asia and the Pacific, New York.

World Bank (2009) World development report 2009: reshaping economic geography, The World Bank, Washington, DC.

World Bank (2012) The rise of metropolitan regions: towards inclusive and sustainable regional development, The World Bank, Jakarta.

World Urbanization Prospects (2011) World urbanization prospects, the 2011 revision, data on cities and urban agglomerations, Retrieved 22 January 2014, from United Nations Department of Economic and Social Affairs http://esa.un.org/unup/CD-ROM/Urban-Agglomerations.htm

Yiftachel, O. (1998) Planning and social control: exploring the dark side, *Journal of Planning Literature*, 12(4), 395–406.