

## **What is to be expected from broad-scale population data? Showcase accessibility model validation using high-resolution census information**

C. Aubrecht<sup>1\*</sup>, G. Yetman<sup>2</sup>, D. Balk<sup>3</sup>, K. Steinnocher<sup>1</sup>

<sup>1\*</sup> AIT Austrian Institute of Technology; Vienna, Austria ([christoph.aubrecht@ait.ac.at](mailto:christoph.aubrecht@ait.ac.at))

<sup>2</sup> Center for International Earth Science Information Network (CIESIN) and  
NASA Socioeconomic Data and Applications Center (SEDAC),  
The Earth Institute at Columbia University; Palisades, NY, USA

<sup>3</sup> Baruch College, City University of New York (CUNY); New York, NY, USA

### **ABSTRACT**

In this paper we give an overview of existing broad-scale population distribution modeling concepts in terms of their different levels of spatial and thematic complexity. The corresponding global or continental-scale datasets are described in terms of their individual characteristics. Special focus is put on using the accessibility concept for reallocating people within spatial reference units based on favoring conditions. This concept is based on the premise that people tend to live in or close to cities and tend to move towards areas that are well connected with urban centers. One example of a continental-scale model is highlighted in that context (African Population Database). Finally, we present a case study where an accessibility surface was produced for Austria, subsequently used for population disaggregation. Referring to available high-resolution census data first validation results provide an idea of what can be expected from super-regionally modeled population datasets.

### **INTRODUCTION**

Dealing with the use of global population data for health applications Balk et al. (2006) presented a list of broad-scale approaches to population distribution modeling, ordered by increasing complexity. The focus of all datasets in this context lies on representing resident population based on the highest resolution input data available.

The first and least complex dataset is the Gridded Population of the World (CIESIN & CIAT, 2004) demonstrating the spatial distribution of human populations across the globe. "The Global Demography Project" has to be cited as first major effort (Tobler et al. 1997) in that context. Produced at the National Center for Geographic Information Analysis (NCGIA) in 1995 this first version of GPW was sort of a precursor for the subsequent series of approaches to consistently map population on a global scale. The current version GPWv3 is the third edition of this broad-scale data product having the purpose to provide a spatially disaggregated population layer that is compatible with datasets from social, economic, and earth science fields. The output shows the distribution of human population converted from national or sub-national spatial units (usually administrative units) of varying resolutions, to a series of geo-referenced quadrilateral grids at a resolution of 2.5 arc minutes (Deichmann et al., 2001). GPWv3 incorporates a number of improvements to the two prior iterations of GPW. Input administrative data have been improved for nearly all of the 232 countries included in the dataset (i.e. the number of administrative units has increased three-fold since GPWv2 and twenty-fold since GPWv1). Additionally, the input data years have been updated for over two-thirds of the countries. The population data estimates, previously only available for 1990 and 1995, are now extended for the period 1990-2015, by quinquennial years. The population estimates for 2005, 2010, and 2015 were produced in collaboration with the United Nations Food and Agriculture Programme (FAO) as 'GPW: Future Estimates' (Balk et al., 2005a).

The Global Rural-Urban Mapping Project (GRUMP)<sup>1</sup> provides a new suite of data products that add rural-urban specification to GPWv3, hence featuring an increased level of complexity (Balk et al., 2005b). This project was developed out of a need for researchers (Montgomery et al., 2003) to be able to distinguish population spatially by urban and rural areas. There is however no single definition of what makes an area “urban”. Balk (2009) points out the United Nations World Urbanization Prospects (UN, 2006) which identifies each country’s definition of the term “urban” whereas criteria include a variety of population size or density thresholds associated with administrative areas, capital cities, and combinations thereof. The central data product resulting from GRUMP is a ‘Gridded Population of the World with Urban Reallocation’ in which spatial and population data of both administrative units and urban extents are gridded at a resolution of 30 arc-seconds (i.e. 25-fold higher resolution compared to GPW). The GRUMP data currently available on the NASA SEDAC website constitute the alpha editions of these products (GRUMP alpha), GRUMP beta which has not yet been released publicly is in the testing phase.

Considering a set of additional assumptions about real-world population distribution the final group of modeled datasets is based on the accessibility concept. Basic motive for that kind of methodology is that people tend to live in or close to cities and tend to move towards areas that are well connected with urban centers (Balk et al., 2006). This premise basically holds true even for rural regions for which it is expected that areas of higher population density are located preferentially close to transport links and bigger cities. The concept of accessibility and related indicators has been in use for a long time in particular in transportation research (Koenig 1980, Halden 2003, Gutiérrez et al. 2010). Applying accessibility indicators to population distribution and reallocation models has however not been that popular until first implementations for producing continental-scale databases for Africa, Asia and Latin America in the mid 1990’s (Deichmann, 1996, 1998). Just recently Langford et al. (2008) examined how alternative population distribution models (i.e. approaches using even-distribution within census tracts vs. dasymetric mapping) influence GIS-based accessibility analyses, thus approaching this topic from the opposite side. Deichmann (1997) gives a comprehensive overview on the use of various accessibility indicators in GIS. He emphasizes that there are several ways to define accessibility and presents different concepts. Accessibility can be defined as the ability for interaction or contact with sites of economic or social opportunity. This definition is supported by Goodall (1987) who states “The concept expresses the ease with which a location may be reached from other locations. [It] summarizes relative opportunities for contact and interaction.”

Besides the abovementioned global and continental-scale population databases there is another dataset available on that spatial level - LandScan<sup>2</sup> - comprising a worldwide population database compiled on a 30 arc-seconds latitude/longitude grid. In addition to transportation networks and populated places LandScan includes likelihood coefficients based on parameters such as elevation, slope, nighttime lights and land cover for apportioning census counts to each grid cell, while less effort is spent on using the highest-possible resolution population input information. This dataset has a categorically different focus to the previously described population distribution models as it aims at measuring ambient population instead of attempting to represent nighttime census resident population (Dobson, 2000; Bhaduri et al., 2002). So, these datasets have different objectives, they make different assumptions, use different methodologies, and are designed to measure two different indicators. As stated on the iSciences blog site (<http://geoserver.isciences.com/DataBlog/>) “it would not be fair to say that one or the other is ‘better’, it is more a question of what tool(s) are best for the job at hand”.

---

<sup>1</sup> GPWv3 and GRUMP data and additional information are provided on the NASA Socioeconomic Data & Applications Center website (<http://sedac.ciesin.columbia.edu/gpw/>, last accessed 25 March 2010).

<sup>2</sup> LandScan data and additional information is provided on the Oak Ridge National Laboratory (ORNL) website (<http://www.ornl.gov/sci/landscan/>, last accessed 25 March 2010).

### ACCESSIBILITY MODELING SAMPLE: AFRICAN POPULATION DATABASE

Serving as example for broad-scale population data based on accessibility modeling the African Population Database will be described in more detail. The currently available dataset is the fourth version of a database of administrative units with associated population figures for Africa (Nelson, 2004). The first version was initially compiled for UNEP's Global Desertification Atlas (UNEP 1992; Deichmann & Eklundh, 1991), with the second and third versions representing corresponding updates and expansions (Deichmann, 1998; WRI, 1995). The newest database release for Africa provides considerably more detail than its three precursors: more than 109,000 administrative units (83,000 of which are in South Africa), compared to about 800 in the first, 2,200 in the second and 4,700 in the third version. In addition, for each of these units a population estimate was compiled for 1960, 70, 80, 90 and 2000 providing an indication of past population dynamics in Africa (figure 1).

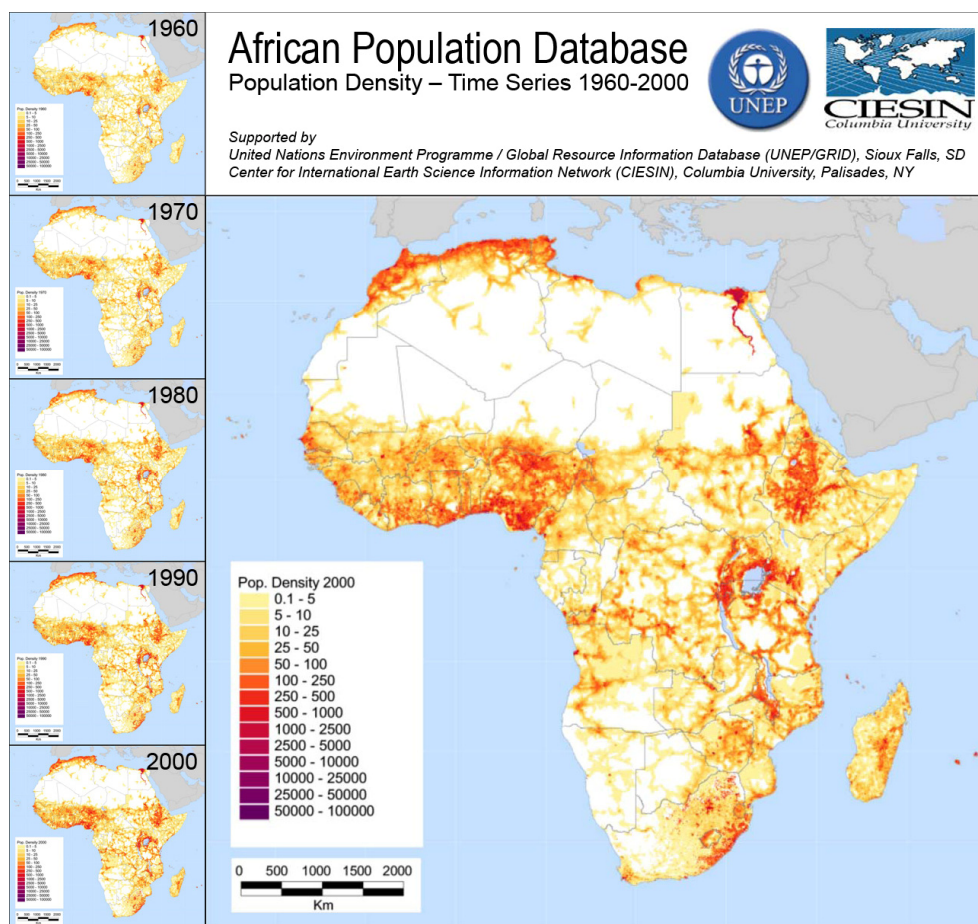


Figure 1: African Population Database - Time Series 1960-2000.

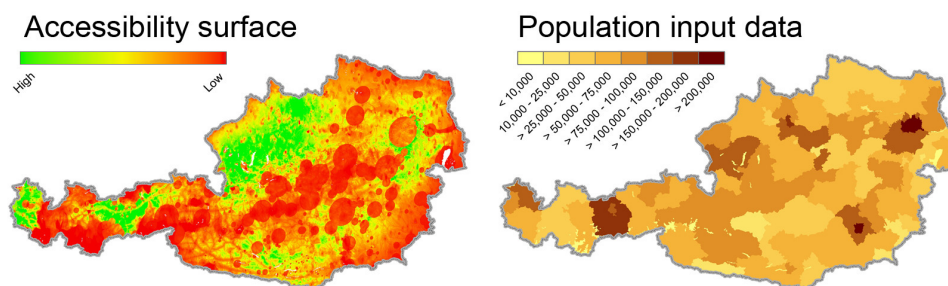
The African administrative boundaries and population database was compiled from a large number of heterogeneous sources. The objective was to create a comprehensive database from existing sources and in a fairly short time period that is suitable for regional or continental-scale applications. The resources available did not allow for in-country data collection or collaboration with national census bureaus, as was done, for example, in the WALTPS study (Brunner et al. 1995). With few exceptions, the datasets thus do not originate from the countries themselves, and input boundary data have not been officially checked or endorsed by the national statistical or mapping agencies.

The method for the development of population raster grids consists of a set of processing steps. The most important input into the model is (1) information about the transportation network consisting of roads, railroads and navigable rivers. The second main component is (2) information on urban centers. Data on the location and size of as many towns and cities as could be identified were collected, and these settlements are linked to the transport network. This information was then used to compute a simple measure of accessibility for each node in the network. This measure, called population potential, is the sum of the population of towns in the vicinity of a specific node weighted by a distance function, where network distances are used rather than straight-line distances. The computed accessibility estimates for each node are subsequently interpolated onto a regular raster surface. Raster data on inland water bodies (lakes and glaciers), protected areas and altitude are then used to adjust the accessibility surface heuristically. Finally, the population totals estimated for each administrative unit are distributed in proportion to the accessibility index measures estimated for each grid cell. The resulting population counts in each pixel could then be converted to densities for further analysis and mapping.

## CASE STUDY: AUSTRIA

### Modeling population distribution

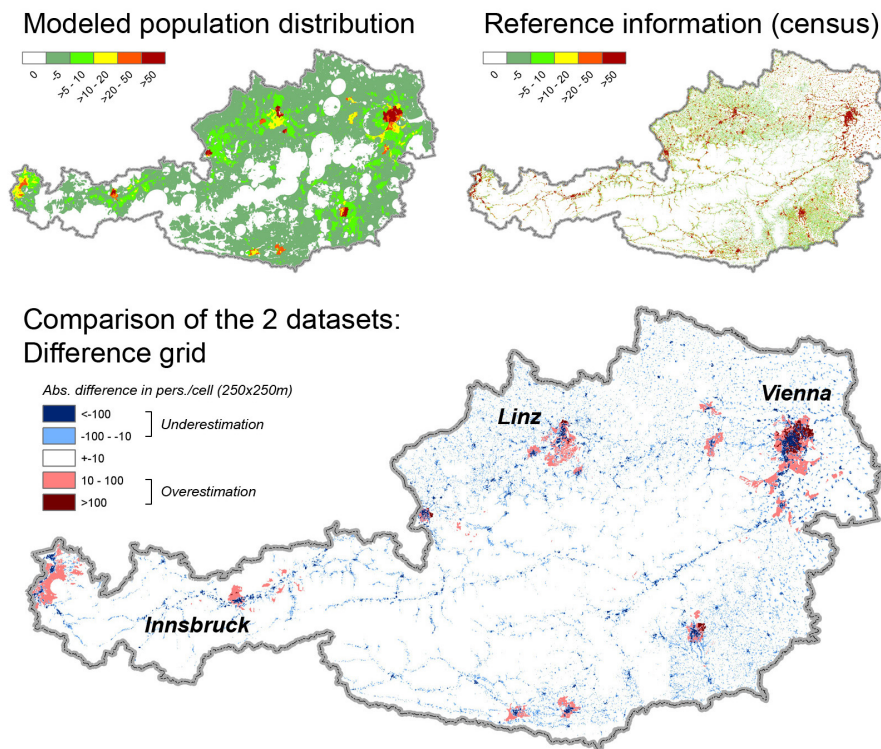
In the following paragraphs we describe how very high resolution population information available from the Austria census 2001 (i.e. 250 m grid cell size) was used as reference to identify and show what can be expected from modeled super-regional and global population datasets. The algorithm (stand-alone program written in the C programming language) that had been used for generation of the African Population Database was adopted for Austria. An accessibility surface was produced (see figure 2) using populated places (GRUMP), water network, road and rail network (GISCO dataset based on EuroGlobalMap EGM 2.0 and EuroRegionalMap ERM 4.0), elevation information (SRTM) and protected areas (World Database on Protected Areas WDPA). Population numbers available on an aggregated district level (i.e. comparable to available official population counts in African countries) were spatially disaggregated based on the calculated accessibility model resulting in a refined population distribution dataset featuring a calculated 250 m resolution.



**Figure 2:** Input for the population distribution - (1) accessibility surface, (2) district level census data.

### Validation using high-resolution census data

First inter-comparison results of modeled population distribution and reference data (figure 3, top left and top right respectively) show regional patterns of the degree of deviation (overall statistics of the study area: standard deviation 40.5, mean error 3.3). As expected, population numbers for suburban regions are overestimated while for the actual urban centers and areas in the broad vicinity of larger cities underestimations are found (compare figure 3: Vienna, Linz, Innsbruck). The latter is mostly due to missing fine-scale information on the spatial distribution of small settlement regions which results in a 'shifting' of population towards the nearby suburban regions. The urban center underestimations can be explained by favoring factors not being weighted high enough in the accessibility model thus not adequately accounting for the highest centrality characteristics. One idea in this context is to include satellite observed nighttime lighting (Elvidge et al., 1997) as centrality indicator and re-run the accessibility model for the study test site. At local scales the population surface has a mottled appearance due to the various input factors. This, and the underestimation of population in urban centers may be aided by applying pycnophylactic interpolation, a mass-conserving algorithm for smoothing data (Tobler, 1979). A problem which becomes obvious in visualizing the data is the inconsistent quality of the WDPA dataset. For Austria no protected area delineations are available, just center-points and corresponding total area information. As protected areas are introduced as limiting factor for population reallocation in the accessibility modeling process, no population is assigned to these areas (see the white circular 'holes' in the modeled population data as shown in figure 3). In Austria, just very few people actually live in formal protected areas, hence no big errors are introduced in the model that way. However, for other countries (e.g. in Africa) that issue can become a major problem.



**Figure 3:** Comparison of modeled population distribution and reference data from the census.



## CONCLUSION

In this paper we gave an overview of existing broad-scale population distribution modeling concepts in terms of their different levels of spatial (i.e. geometric resolution) and thematic (i.e. input information) complexity. Characteristics of the individual corresponding global or continental-scale datasets are specified starting with the current version 3 release of the 'Gridded Population of the World (GPW)' database, followed by the 'Global Rural-Urban Mapping Project (GRUMP)' which considers rural-urban discrepancies for census population disaggregation. The next level of complexity for population distribution modeling was examined in more detail. Using the accessibility concept widely used in transportation research (i.e. based on the premise that people tend to live in or close to cities and tend to move towards areas that are well connected with urban centers) people are reallocated within spatial reference units based on favoring conditions, implemented in the modeling process by using weighting factors. The 'African Population Database' was introduced as an example for a continental-scale model. Finally, we presented a case study where an accessibility surface had been produced for Austria, subsequently used for population disaggregation. Referring to available high-resolution census data first validation results provide an idea of what can be expected from super-regionally modeled population datasets. Providing a good first impression of population distribution patterns, the main problems are related to urban centers and their immediate vicinity. For future approaches we therefore consider including a centrality indicator based on satellite observed nighttime lighting data.

## BIBLIOGRAPHY

- Balk, D., 2009, More than a name: Why is Global Urban Population Mapping a GRUMPY proposition? In Gamba, P., and Herold, M. (eds.) *Global Mapping of Human Settlement: Experiences, Data Sets, and Prospects*. Taylor & Francis Series in Remote Sensing Applications, CRC Press.
- Balk, D., Brickmann, M., Anderson, B., Pozzi, F., Yetman, G., 2005a, Mapping global urban and rural population distributions. Annex Estimates of future global population distribution to 2015. Environmental and natural resources working paper 24, FAO, Rome, Italy.
- Balk, D., Pozzi, F., Yetman, G., Deichmann, U., and Nelson, A., 2005b, The distribution of people and the dimension of place: Methodologies to improve the global estimation of urban extents. Proceedings, 5th International Symposium Remote Sensing of Urban Areas (URS 2005), Tempe, AZ, USA, March 2005.
- Balk, D.L., Deichmann, U., Yetman, G., Pozzi, F., Hay, S.I., and Nelson, A., 2006, Determining Global Population Distribution: Methods, Applications and Data. In Hay, S.I., Graham, A.J. and Rogers, D.J. (eds) *Global mapping of infectious diseases: Methods, Examples and Emerging Applications*. Advances in Parasitology, vol. 62. London: Academic Press. 119-156.
- Bhaduri, B., Bright, E., Coleman, P., and Dobson, J., 2002, LandScan: Locating People is What Matters. *Geoinformatics* 5(2): 34-37.
- Brunner, J., Henninger, N., Deichmann, U., and Ninnin, B., 1995, West Africa Long Term Perspective Study (WALTPS) - Database and User's Guide, World Resources Institute, Washington DC, in collaboration with the Club du Sahel, Paris.
- Center for International Earth Science Information Network (CIESIN), and Centro Internacional de Agricultura Tropical (CIAT), 2004, Gridded Population of the World, version 3 (GPWv3) and the Global Rural-Urban Mapping Project (GRUMP). Palisades, NY: Columbia University. Available at <http://beta.sedac.ciesin.columbia.edu/gpw> (last accessed 25 March 2010).
- Deichmann, U., 1996, Asia medium resolution population database documentation, NCGIA, UC Santa Barbara and UNEP/GRID Geneva (<http://grid2.cr.usgs.gov/globalpop/>, last accessed 25 March 2010).

- Deichmann, U., 1997, Accessibility indicators in GIS. United Nations Statistics Division, Department for Economic and Policy Analysis, New York, NY, USA.
- Deichmann, U., 1998, Africa medium resolution population database documentation, NCGIA, UC Santa Barbara, UNEP/GRID Sioux Falls and World Resources Institute (<http://grid2.cr.usgs.gov/globalpop/>, last accessed 25 March 2010).
- Deichmann, U., and Eklundh, L., 1991, Global digital datasets for land degradation studies: A GIS approach, GRID Case Study Series No. 4, Global Resource Information Database, United Nations Environment Programme, Nairobi.
- Deichmann, U., Balk, D., Yetman, G., 2001, Transforming Population Data for Interdisciplinary Usages: From census to grid. GPWv2 working paper. CIESIN, Columbia University, Palisades, NY, USA.
- Dobson, J.E., Bright, E.A., Coleman, P.R., Durfee, R.C., and Worley, B.A., 2000, LandScan: A Global Population Database for Estimating Populations at Risk. *Photogrammetric Engineering and Remote Sensing* 66(7): 849-857.
- Elvidge, C.D., Baugh, K.E., Kihn, E.A., Kroehl, H.W., and Davis, E.R., 1997, Mapping of city lights using DMSP Operational Linescan System data. *Photogrammetric Engineering and Remote Sensing* 63: 727-734.
- Goodall, B., 1987, *Dictionary of human geography*. Harmondsworth, Penguin.
- Gutiérrez, J., Condeço-Melhorado, A., and Martín, J.C., 2010, Using accessibility indicators and GIS to assess spatial spillovers of transport infrastructure investment. *Journal of Transport Geography* 18(1): 141-152.
- Halden, D., 2003, Accessibility Analysis Concepts and their Application to Transport Policy, Programme and Project Evaluation. In A. Pearman, P. Mackie, and J. Nellthorp (eds.) *Transport Projects, Programmes and Policies: Evaluation Needs and Capabilities*. Ashgate.
- Koenig, J.G., 1980, Indicators of urban accessibility: Theory and application. *Transportation* 9(2): 145-172.
- Langford, M., Higgs, G., Radcliffe, J., and White, S., 2008, Urban population distribution models and service accessibility estimation. *Computers, Environment and Urban Systems* 32(1): 66-80.
- Montgomery, M.R., Stern, R., Cohen, B., and Reed, H., 2003, *Cities Transformed: Demographic Change and Its Implications in the Developing World*. Washington, DC: NRC.
- Nelson, A., 2004, African Population Database. New version. Digital database and documentation. Sioux Falls, SD: UNEP-GRID Global Resource Information Database.
- Tobler, W., 1979, Smooth Pycnophylactic Interpolation for Geographical Regions. *Journal of the American Statistical Association* 74(367): 519-530.
- Tobler, W., Deichmann, U., Gottsegen, J., and Maloy, K., 1997, World Population in a Grid of Spherical Quadrilaterals. *International Journal of Population Geography* 3: 203-225.
- United Nations Environment Programme (UNEP), 1992, *World Atlas of Desertification*, Edward Arnold Publishers, London.
- United Nations, 2006, World population prospects: the 2006 revision. Newest releases available at <http://esa.un.org/unpp> and <http://unstats.un.org> (last accessed 25 March 2010).
- WRI, 1995, Africa data sampler, Digital database and documentation, World Resources Institute, Washington, D.C.