Quantifying spatial uncertainty of population estimates: Evidence from three Tanzanian districts

Kevin M. Mwenda[*], Phaedon C. Kyriakidis[**], David López-Carr[*].

Abstract

This paper explores how population estimates in three Tanzanian districts might vary with the utilization of fine-scale population estimates. Preliminary research suggests that the use of coarse aggregated values such as census data at the district level instead of spatially varying values may hinder effective decision analysis in situations that call for the use of a spatial population dataset. We show, for example, that the number of adults of reproductive age (15-49) is sensitive to the level of aggregation of the population datasets and as such, may result in a significantly varying results. Decision-makers may benefit from the application of multiple realizations of gridded population instead of using coarse values. In so doing, one could potentially mitigate unforeseeable consequences of underestimating or overestimating population-related impact assessment outcomes in developing countries such as Tanzania.

Keywords

WorldPop, Poisson, Simulation, Uncertainty.

[*] University of California, Santa Barbara, USA

[**] Cyprus University of Technology, Limassol, Cyprus

Introduction

In most developing countries, the low resolution aggregated census data that are used to create the high resolution, disaggregated or downscaled population datasets show significant variation by year and by spatial resolution (Linard and Tatem, 2012). In addition, contemporary census data are not usually available for such countries and they therefore typically rely on census data that is around a decade old and collected at coarse administrative units (Tatem *et al.*, 2011). The cumulative effect of using varying years of census data, aggregated intercensal growth rates and adjustments to estimate total population results in downscaled datasets have significant variations in estimated population size and spatial distribution (Tatem *et al.* 2011). The choice of such datasets that 'hide' uncertainty behind their input data, methods, and output estimates, may lead to significant discrepancies in vulnerability studies by decision-makers (e.g. researchers, policy-makers and/or agencies) who use the data (Thompson and Graham, 1996).

This study explores how the adults of reproductive age (15 – 49 years of age, henceforth refered to as adult population) estimates in three districts in Tanzania might change with the utilization of realizations of gridded estimates.

Study area

Tanzania is situated in Eastern Africa between Longitude 29° and 41° East and Latitude 1° and 12° South. The country consists of a mainland and set of islands with 169 districts and a total land area of nearly 900,000 square kilometres (United Republic of Tanzania 2012). Three districts of varying geographic and demographic characteristics were selected for a proof of concept under diverse population and areal contexts. The first district is

Kinondoni; it is the largest of the three with an area of 537 km^2 and it experienced the most population gain in the country between 2010 and 2012 (~630,000 people). The second district is *Mbeya Urban* district which is roughly half the size of *Kinondoni* district but had the least population change in Tanzania between 2010 and 2012 (~500 people). The third one is *Mjini* district which is the smallest district in Tanzania (15 km^2) but has the highest population density of the three districts.

Datasets

The Tanzania 2012 census data together with the geographic boundary files were obtained from the National Bureau of Statistics (NBS) website, after which the adult population data were calculated at the district-level (Tanzania National Bureau of Statistics 2013). Gridded population data for Tanzania from 2010 was obtained from WorldPop (formerly AfriPop project) at a resolution of 100m (Tatem *et al.*, 2007).

Methods

First, gridded population estimates in 2010 were used as a spatial pattern surrogate for gridded population estimates in 2012. Let $\hat{x}_i$ represent population gridded at the i^{th} pixel contained in district k, where i=1,…,N_k pixels. A population change factor δ_k depicting a change of aggregate population from 2010 to 2012 was calculated using the following equation: $\delta_k = \frac{\bar{y}_k}{\bar{\hat{x}}_k}$, where $\bar{\hat{x}}_k$ represents aggregate population estimates in each district k in 2010 and is calculated as: $\bar{\hat{x}}_k = \sum_{i=1}^{N_k} \hat{x}_i$. The population change factor

δ_k was then applied to gridded estimates $\hat{x}_i$ to obtain gridded population 2012 estimates $\hat{y}_i$ as follows:

$$\hat{y}_i = \delta_k * \hat{x}_i.$$

Finally, the 2012 gridded adult population estimates $\hat{z}_i$ were obtained using the equation: $\hat{z}_i = f_k * \hat{y}_i$.

We then employed the non-homogenous Poisson process to simulate multiple realizations of the gridded adult population z_i since the latter was considered an outcome of a counting process depicting a population that occurred in a particular space and time. The gridded adult population realizations $(\bar{z}_k)$ were adjusted in order to match the spatial pattern of the WorldPop dataset while summing up to the estimated census 'target' adult population $\bar{z}_k{}^t$ that was previously definedas the aggregate census population $\bar{z}_k$ for each district *k*. We developed a 'spatial matching' algorithm to adjust realizations in which $\bar{z}_k \neq \bar{z}_k{}^t$. For realizations in which $\bar{z}_k$>$\bar{z}_k{}^t$, a threshold ϕ was selected as the median of z_i for which changes were only be made randomly to pixels for which $z_i > \phi$.

Consequently, any randomly selected pixels had their respective populations decreased by a randomly selected net migration rate ϖ where $m_l \leq \varpi \leq m_h$, until $\bar{z}_k = \bar{z}_k{}^t$. The net migration rate m_l represents a "lower boundary" national net emigration rate of 0.29 migrants per 1000 people as estimated for Tanzania in 2012 (CIA, 2015) and the net migration rate$m_h$represents a "high boundary" national net emigration rate of 0.63% as estimated in 2013 (IOM, 2015). Each randomly selected net migration rate ϖ was converted into a ratio on a 0-1 scale and used to decrease z_i by a factor of $(1 - \varpi)$ as follows: $z_i{}' = (1 - \varpi) * (z_i)$. In the aforementioned equation, $z_i{}'$ is the adjusted simulation of gridded adult populationz_i. During the adjustment process, adjusted simulated values $z_i{}'$ replaced the original z_i, and the adjustment was considered complete when $\sum_{i=1}^{N_k} z_i = \bar{z}_k = \bar{z}_k{}^t$. The algorithm then checked if there was any adjusted realization in which $\sum_{i=1}^{N_k} z_i < \bar{z}_k{}^t$ Where found, the algorithm increased $\sum_{i=1}^{N_k} z_i$by $[\bar{z}_k{}^t - \bar{z}_k]$ to match $\bar{z}_k{}^t$. For realizations in which $\bar{z}_k$<$\bar{z}_k{}^t$, a similar threshold ϕ was selected as the median of z_i for which changes were only made randomly to pixels in which z_i>ϕ_i. Consequently, any randomly selected pixels had their respective populations increased by a randomly selected urbanization rate ψ where $b_l \leq \psi \leq b_h$ until $\bar{z}_k = \bar{z}_k{}^t$. The urbanization rate b_l represents a "low boundary" annual urbanization rate of 4.2%, estimated for Tanzania for the period between 2005-2010 (UN 2015) and the urbanization rate b_h represents a "high boundary" annual urbanization rate of 4.77%, estimated for the period between 2010-2015 (CIA, 2015). Each randomly selected urbanization rate ψ was converted into a ratio on a 0-1 scale and used to increase z_i as follows: $z_i{}' = z_i + (\psi * z_i)$. The adjustment was considered complete when $\sum_{i=1}^{N_k} z_i = \bar{z}_k = \bar{z}_k{}^t$. The algorithm then checked if there was any adjusted realization in which $\sum_{i=1}^{N_k} z_i > \bar{z}_k{}^t$. Where found, the algorithm decreased $\sum_{i=1}^{N_k} z_i$by $[\bar{z}_k{}^t - \bar{z}_k]$ to match $\bar{z}_k{}^t$.

Lastly, variances of $\sum_{i=1}^{N_k} z_i$ are explicated by calculating confidence intervals around $\bar{z}_k{}^t$, based on multiple realizations of gridded estimates.

Preliminary Results

Preliminary findings show that adult population estimates are sensitive to the level of aggregation of the population datasets. The adult population coarse values for the three districts were as follows – *Kinondoni* district: 1,106,940 (95% confidence interval [CI]: 1,106,670; 1,107,100):

Mbeya Urban district: 140,922 (95% CI: 140,880; 141,030) and *Mjini* district: 120,585 (95% CI: 120,570; 120,710).

Conclusion

Ourfindings show that assessment of population values is sensitive to the level of aggregation of the population datasets and as such, may result in a significantly varying number of adult population, a discrepancy that could potentially have far-reaching impacts on population-dependent policies on a national level. However, we recognize that up-to-date fine resolution datasets for some regions, especially in developing countries, are not always readily available. Furthermore, the few that exist typically contain source data and methods that are inherently uncertain and difficult to explicate. In such cases, decision-makers may benefit from the application of multiple simulated spatial distributions of fine scale population along with the associated impact values. From this distribution of impact estimates, decision-makers should feel relatively confident selecting optimized values based on an explication of risk attitude, thus avoiding unforeseeable consequences of underestimating or overestimating impact assessment outcomes.

References

Central Intelligence Agency (CIA) (2015), *The World Factbook: Tanzania* [https://www.cia.gov/library/publications/the-world-factbook/geos/tz.html]

International Organization for Migration (IOM) (2015), *World Migration,* [https://www.iom.int/world-migration].

Linard C., Tatem A. (2012), Large-scale spatial population databases in infectious disease research, *International Journal of Health Geographics,* 11(7).

Tanzania National Bureau of Statistics (2013), *Statistics for Development* [http://www.nbs.go.tz/].

Tatem A., Noor A., von Hagen C., Di Gregorio A., Hay S. (2007), High resolution population maps for low income nations: Combining land cover and census in East Africa, *PlOs ONE,* 2(12).

Tatem A., Campiz N., Gething P., Snow R., Linard C. (2011), The effects of spatial population dataset choice on estimated population at risk of disease, *Population Health Metrics,* 9(4).

Thompson K., Graham J. (1996), Going beyond the single number: Using probabilistic risk assessment to improve risk management, *Human Ecological Risk Assessment:* 2(4), pp. 1008–1034

United Nations (UN) (2015), *Population Division* [http://www.un.org/en/development/desa/population]

United Republic of Tanzania (2012), *Tanzania in Figures 2012* [http://www.nbs.go.tz/nbs/takwimu/references/Tanzania_in_figures2012.pdf].