GeoJournal (2011) 76:525–538
DOI 10.1007/s10708-010-9364-8

Assessing the use of global land cover data for guiding large area population distribution modelling

Catherine Linard · Marius Gilbert · Andrew J. Tatem

Published online: 25 May 2010

Abstract Gridded population distribution data are finding increasing use in a wide range of fields, including resource allocation, disease burden estimation and climate change impact assessment. Land cover information can be used in combination with detailed settlement extents to redistribute aggregated census counts to improve the accuracy of national-scale gridded population data. In East Africa, such analyses have been done using regional land cover data, thus restricting application of the approach to this region. If gridded population data are to be improved across Africa, an alternative, consistent and comparable source of land cover data is required. Here these analyses were repeated for Kenya using four continent-wide land cover datasets combined with detailed settlement extents and accuracies were assessed against detailed census data. The aim was to identify the large area land cover dataset that, combined with detailed settlement extents, produce the most accurate population distribution data. The effectiveness of the population distribution modelling procedures in the absence of high resolution census data was evaluated, as was the extrapolation ability of population densities between different regions. Results showed that the use of the GlobCover dataset refined with detailed settlement extents provided significantly more accurate gridded population data compared to the use of refined AVHRR-derived, MODIS-derived and GLC2000 land cover datasets. This study supports the hypothesis that land cover information is important for improving population distribution model accuracies, particularly in countries where only coarse resolution census data are available. Obtaining high resolution census data must however remain the priority. With its higher spatial resolution and its more recent data acquisition, the GlobCover dataset was found as the most valuable resource to use in combination with detailed settlement extents for the production of gridded population datasets across large areas.

Keywords Population mapping · Global land cover data · Census data · Dasymetric modelling · GlobCover

Electronic supplementary material The online version of this article (doi:10.1007/s10708-010-9364-8) contains supplementary material, which is available to authorized users.

C. Linard (✉)
Spatial Ecology and Epidemiology Group, Department of Zoology, University of Oxford, South Parks Road, OX1 3PS Oxford, UK
e-mail: catherine.linard@zoo.ox.ac.uk

M. Gilbert
Biological Control and Spatial Ecology, Université Libre de Bruxelles, CP 160/12, 50, Avenue F.D. Roosevelt 50, 1050 Brussels, Belgium

A. J. Tatem
Emerging Pathogens Institute and Department of Geography, University of Florida, Gainesville, FL 32611-7315, USA

Introduction

Gridded population distribution data are increasingly being used for resource allocation, disease burden estimation and climate change impact assessment amongst other applications, at global, continental and national scales. Detailed and spatially disaggregated population data are essential resources in the assessment of the number of impacted people in decision-making processes related to developmental or health issues (Bhaduri et al. 2002; Dobson et al. 2000; Hay et al. 2005; Salvatore et al. 2005). Existing gridded population data have been used, for example, to quantify populations at risk of several infectious diseases such as malaria (Guerra et al. 2006; Hay et al. 2009), yellow fever and dengue (Rogers et al. 2006), or avian influenza (Ferguson et al. 2005; Rao et al. 2009). Global population datasets have also been used to study the spatial distribution of infant mortality (Storeygard et al. 2008) and child hunger (Balk et al. 2005c). Moreover, gridded population distribution data have shown application in the analysis of the impacts of climate change, such as sea level rise (McGranahan et al. 2007) and the collapse of an Antarctic ice sheet (Nicholls et al. 2005), while the vulnerability of people to natural disasters has also been quantified (Balk et al. 2005a; Maynard-Ford et al. 2008).

Three global gridded population datasets are available for undertaking such studies; the Gridded Population of the World (GPW), the Global Rural Urban Mapping Project (GRUMP), and the LandScan Global Population database. The United Nation Environment Programme (UNEP) has also compiled gridded population data for Africa, Asia and Latin America. In the GPW database–which was first released in 1995 (Tobler et al. 1995, 1997), then updated in 2000 (Deichmann et al. 2001) and 2004 (Balk and Yetman 2004)–population data were simply areal-weighted per administrative unit, thus assuming that the population is uniformly distributed within each administrative unit. GRUMP uses a similar approach to GPW, but incorporates satellite nighttime light-derived urban extents and their corresponding populations in the spatial reallocation of census counts (Balk et al. 2005b). LandScan was first developed in 1998 (Dobson et al. 2000), then updated yearly from 2000 to 2008. LandScan uses ancillary data such as roads, slope, land cover and nighttime lights to estimate probabilities of population occurrence in grid cells. Populations are spatially reallocated within each areal unit using modelling approaches based on these probability coefficients (Dobson et al. 2000; Bhaduri et al. 2007). Finally, the UNEP database was constructed based on an accessibility surface developed from road networks and populated places datasets (Deichmann 1996; Hyman et al. 2004; Nelson 2004).

These existing large area population datasets exhibit significant drawbacks due to the coarse nature of the input census data used in their construction for many countries, particularly those in the low income regions of the World. For the majority of African countries, census data are often over a decade old and at a provincial or district level resolution (Tatem et al. 2008). The use of modelling techniques for the spatial reallocation of populations within census units is therefore particularly relevant for Africa. Dasymetric modelling methods involve using ancillary data to redistribute populations from administrative units to more homogenous units such as square grids (Mennis 2003). However, these approaches only increase population distribution model accuracies over the simple gridding (areal weighting) of census data if the ancillary data is more detailed and complete spatially than the input census data, and can be detrimental to modelling accuracies otherwise (Hay et al. 2005; Tatem et al. 2007). Land cover and land use data, particularly on settlements, at a spatial resolution finer than the scale of census data administrative units offer an opportunity for improving population distribution models in areas with poor ancillary spatial data, such as sub-Saharan Africa. Population density is assumed to vary according to land use and land cover types (Mennis 2003; Wright 1936). Land use classes–defined by purposes for which humans exploit the land cover–are closely linked to people activities, which make it a more effective indicator of population distribution than land cover. Satellite remote sensing offers a cheap and effective solution to obtain spatial information such as land cover and land use data at different spatial scales (Tatem et al. 2004).

Recent work forming part of the AfriPop project (www.afripop.org) has shown that detailed satellite imagery-based mapping of settlements combined with land cover information can be used to increase population model accuracies across large areas (Tatem et al. 2007). Using East Africa as an example, Tatem et al. (2007) showed that the combination of detailed settlement extents data with land cover data produced more accurate population distribution data than simple areal weighting or the allocation of people only to the grid squares classified as settlement. Dasymetric modelling methods based on land use data require the definition of relative weights associated with land use classes (Hay et al. 2005; Tatem et al. 2007). These weights are first calculated for regions where high resolution census data are available and then applied to other geographically proximate or similar regions with coarser census data. The aim of the AfriPop project is to extend these dasymetric methods to model population distributions across the whole of Africa. As census data are coarse and outdated in many of these countries, land cover specific weights will be calculated based on regions where accurate, detailed and contemporary data are available and then extrapolated to neighbouring regions. The extrapolation level will depend on available data. This spatial extrapolation of relative population weights assumes that the weights are consistent across the regions considered.

The work performed by Tatem et al. (2007) relied upon East Africa-specific land cover information (Africover, www.africover.org), thus restricting application to East Africa. The extension of these approaches beyond the region requires the identification and testing of candidate land cover datasets of wider extent. This paper aims to identify the large area land cover dataset that, combined with detailed settlement extents, produces the most accurate population distribution data. The most appropriate land cover data, refined with detailed settlement extents, will then be used for population distribution modelling across Africa. Here, four satellite imagery derived global land cover datasets are first refined in the same way, and then tested with Kenyan census data on their ability to improve the accuracy of population distribution models. In addition, the spatial extrapolation ability of the relative weights calculated from the four refined land cover datasets was also tested.

Data

Land cover and land use

Four freely available global land cover datasets were acquired. The main characteristics of these four global land cover datasets along with their sources are presented in Table 1. The first one is a global land

Table 1 Global land cover datasets and their main characteristics

Code	Map	Producer	Number of land cover classes	Spatial resolution	Data acquisition year	Source
AVHRR	Advanced Very High Resolution Radiometer (AVHRR) Land Cover Classification	University of Maryland, Department of Geography	14	1 km	1981–1994	http://glcf.umiacs.umd.edu/data/landcover/
MODIS	MODerate resolution Imaging Spectroradiometer (MODIS) Land Cover Classification	Boston University, Department of Geography	19	1 km	2001	http://duckwater.bu.edu/lc/mod12q1.html
GLC2000	Global land cover 2000 (v1.1)	European Commission Joint Research Center	24	1 km	2000	http://bioval.jrc.ec.europa.eu/products/glc2000/glc2000.php
GlobCover	GlobCover Land Cover product v2.2	European Space Agency	23 global and 47 regional	300 m	2005–2006	http://ionia1.esrin.esa.int/

cover classification at a spatial resolution of 1 km, using 14 years of imagery from the NASA/NOAA Pathfinder Land (PAL) Advanced Very High Resolution Radiometer (AVHRR) dataset (Hansen et al. 2000). A second global land cover classification at 1 km spatial resolution was obtained, this time using 1 year of Moderate Resolution Imaging Spectrometer (MODIS) data (Friedl et al. 2002). Thirdly, the Global Land Cover 2000 (GLC2000) dataset was acquired. This 1 km spatial resolution global land cover dataset was derived from daily global images from the VEGETATION sensor on board the SPOT 4 satellite over a 14 month period (Fritz et al. 2002). Finally, the GlobCover Land Cover product (GlobCover) was obtained. This most recent land cover dataset, with a spatial resolution of 300 meters and compatible with the UN Land Cover Classification System (LCCS), was derived from a time-series of Medium Resolution Imaging Spectrometer (MERIS) images acquired from December 2004 to June 2006 (Arino et al. 2007, 2008). These four datasets describe mainly land cover features, but also give some land use information.

Settlements

Settlement maps at 30 m spatial resolution were created by Tatem et al. (2007) for five East African countries (Kenya, Uganda, Burundi, Rwanda and Tanzania) based upon methodologies detailed in Tatem et al. (2004). In brief, bands 1–5, 7 and 8 from Landsat Enhanced Thematic Mapper (ETM) imagery and eight texture layers extracted from Radarsat-1 synthetic aperture radar (SAR) were combined for classifier training. The imagery was split into segments and spatial-spectral segmentation was undertaken in each segment. A feed-forward neural network classifier was then used to identify settlements within each spectrally and spatially contiguous zone, using Africover and settlement centroid data for training and testing. In highly rugged areas, only ETM data were used to avoid strong radar responses due to variations in topography.

Census

Administrative unit level 0 (national), 1 (province), 2 (district), 3 (division), 4 (location), 5 (sublocation) Kenya census data were obtained from the 1999 population and housing census report, available at the Central Bureau of Statistics in Nairobi (CBS 2001), along with corresponding administrative unit boundaries. Also obtained were corresponding census data at administrative unit level 6 (enumeration area) with corresponding boundaries for 58 of the 69 Kenyan districts.

Methods

Population distribution modelling approach

Here we use land cover datasets that cover the world combined with detailed settlement extents and census data to produce gridded population distribution data for Kenya. Four main methodological stages were undertaken: (1) refining of the settlement extents of the global land cover data, (2) dividing enumeration areas in two samples, (3) deriving land cover specific weights and modelling population distribution based on each refined global land cover dataset and (4) assessing the accuracy of the population distribution models produced. Fig. 1 summarizes the whole procedure and shows these four parts in different boxes.

Land cover data refinement (Fig. 1, box 1)

The global land cover maps were 'refined' to accommodate the more detailed and accurate information on settlements provided by Tatem et al. (2007). The four global land cover datasets were first resampled to 100 m spatial resolution. For each land cover dataset, the urban class, which typically overestimates settlement extent size (Tatem et al. 2005, 2007), was removed and the surrounding classes expanded equally to fill the remaining space. The 30 m settlement map constructed in Tatem et al. (2007) was also degraded to 100 m spatial resolution. This more detailed settlement map was then overlaid onto the 'urban class deprived' land cover map and land covers beneath were replaced to produce a refined land cover map. Four refined land cover datasets were therefore created for Kenya.

Sampling methods (Fig. 1, box 2)

In order to use different datasets for modelling and accuracy assessment, the 46,034 Kenyan enumeration areas (EA) were divided in two samples. Different

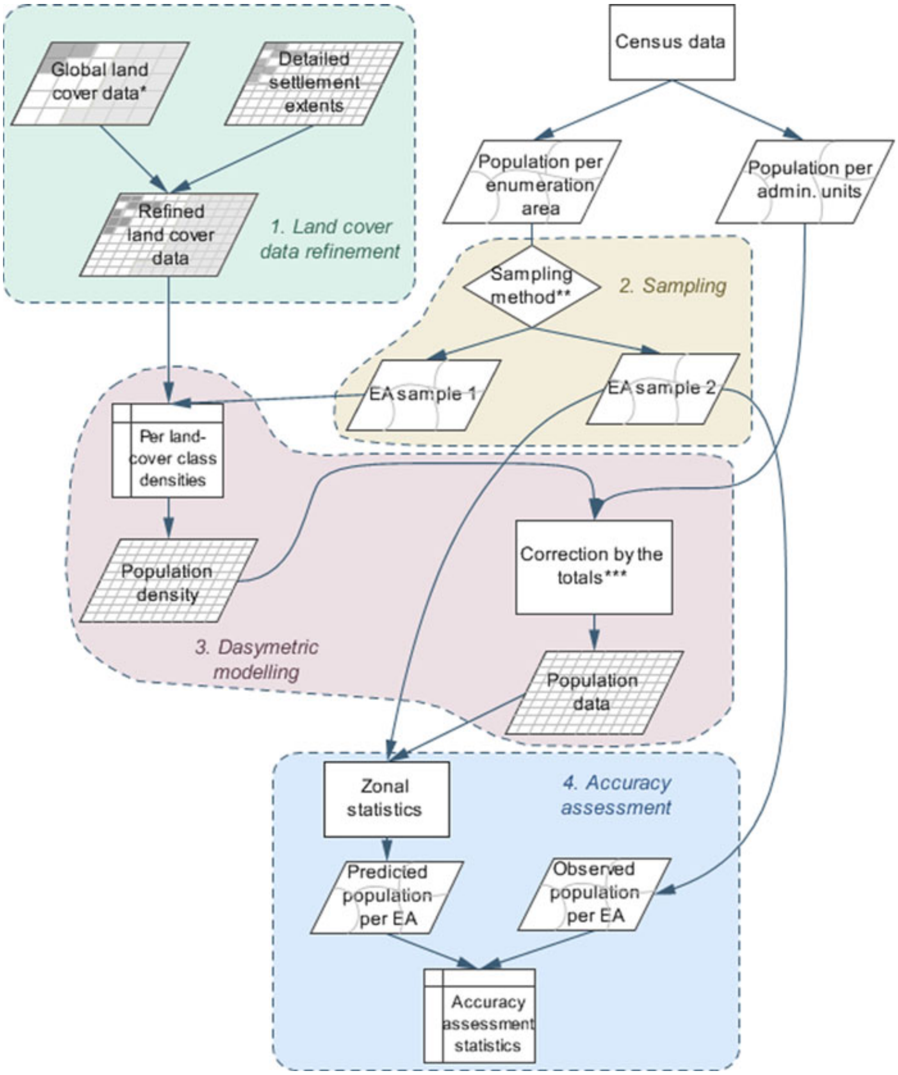
Global land cover data*
Detailed settlement extents
Refined land cover data
1. Land cover data refinement
Census data
Population per enumeration area
Population per admin. units
Sampling method**
2. Sampling
EA sample 1
EA sample 2
Per land-cover class densities
Population density
Correction by the totals***
Population data
3. Dasymetric modelling
Zonal statistics
4. Accuracy assessment
Predicted population per EA
Observed population per EA
Accuracy assessment statistics

Fig. 1 Summary of the methodology followed in this paper. * Global land cover data: AVHRR, MODIS, GLC2000 or GlobCover (Table 1). ** Sampling method: depends on EXL, i.e. RD, L5, L4, L3, L2 or L1 (Table 2). *** Totals correction methods (TCM): depends on the level of administrative data used, i.e. ADMIN-5, ADMIN-4, ADMIN-3, ADMIN-2, ADMIN-1, ADMIN-0 or no correction by the totals (Table 3)

sampling methods were used in order to evaluate the extrapolation ability of the spatial population data production. Here we tested the impact of an increasing extrapolation level (EXL) on the precision of population data produced. The EXL represents the level at which population weights are extrapolated, from close and similar regions to more distant and environmentally different regions. The EXL only determines the sampling method used in the population modelling procedure. With a low EXL, EAs used for modelling and EAs used for accuracy assessment were chosen randomly. With higher EXL levels, EAs were selected based on the administrative unit they belong to: EAs from half of the administrative units were selected for modelling, and the other half was used for accuracy assessment. For example, the sampling method with maximum EXL (EXL = L1) randomly selects 4 of the 8 Kenyan provinces, EAs belonging to the 4 selected provinces constituting the first modelling sample and other EAs constituting the second accuracy assessment sample. In this case, the two samples are much more clustered, and population weights are extrapolated from one province to the other. Table 2 shows the different EXL with their corresponding sampling method.

Dasymetric modelling (Fig. 1, box 3)

The refined land cover data and Kenyan enumeration area census data were then used to define per land

Table 2 Extrapolation levels (EXL) with their corresponding sampling method

EXL	Sampling method
RD	Random selection
L5	Selection based on admin. level 5 (sublocations)
L4	Selection based on admin. level 4 (locations)
L3	Selection based on admin. level 3 (divisions)
L2	Selection based on admin. level 2 (districts)
L1	Selection based on admin. level 1 (provinces)

The sampling method divides EAs in two samples–one for modeling and one for accuracy assessment–in order to evaluate the extrapolation ability of the population modelling approach

Table 3 Totals correction methods (TCM) with their corresponding average spatial resolution (ASR) in Kenya

TCM	Administrative level	ASR (km)
ADMIN-5	Sublocation	9.35
ADMIN-4	Location	15.73
ADMIN-3	Division	35.01
ADMIN-2	District	91.65
ADMIN-1	Province	269.17
ADMIN-0	National	761.33
No correction	/	/

The TCM is the level of administrative data used to correct population numbers by the administrative unit totals. The ASR measures the effective resolution of administrative units in kilometers. It is calculated as the square root of the land area divided by the number of administrative units (Balk and Yetman 2004)

cover class population densities (i.e. the average number of people per 100 × 100 m pixel). Mennis and Hultgren (2006) described and compared different methods for estimating population densities based on land cover data. Here, the average population density of one specific land cover class was calculated based on EAs from the first sample that record this land cover class for the majority of their pixels. Different tables were produced containing the population density per land cover class for each of the four newly created land cover datasets. Zeros were attributed to classes with no human habitation, mainly water bodies.

These per land cover class densities were then used as weights to reallocate populations within Kenyan administrative units. In one administrative unit, the sum of per-pixel population counts is therefore equivalent to the census population data. The administrative unit level of the census data used to adjust population totals is defined by the TCM (totals correction method). The population modelling procedure was repeated using different TCM, i.e. census data at different administrative levels, in order to explore the effectiveness of the population modelling procedures in the absence of high resolution census data (Table 3). To facilitate the comparison with available census data in other countries, the TCM used is associated with the average spatial resolution (ASR) of the administrative unit level of census data in Kenya. The ASR measures the effective resolution of administrative units in kilometres. It is calculated as the square root of the land area divided by the number of administrative units (Balk and Yetman 2004). Different 100 m population distribution datasets were created for the entire of Kenya based on the land cover data and the totals correction method (TCM) used. The gridded population data produced are not projected, but are referenced by geographic WGS84 coordinates.

Accuracy assessment (Fig. 1, box 4)

The accuracies of these population distribution data were tested principally using the second sample of EA census data, the first sample having been used for the relative weights calculation. With an average of 23,017 EAs and an ASR of 3.21 km (8.4 EAs per sublocation in average), these provided a valuable dataset for assessing the accuracy with which populations had been distributed within each administrative unit by the application of each global land cover data. Predicted population data per EA were compared to observed population data from the 1999 Kenyan census. Accuracy statistics including root mean square errors (RMSE) and Pearson correlation coefficients were computed. Accuracies were also tested by comparing the output population distribution data derived from each land cover product to areal weighting, to examine which approaches produced improvements over this simplest of methods. As discussed previously, the areal weighting method is a simple population distribution modelling method consisting of a homogenous distribution of populations within census units, and represents the basis by which the existing widely used global population data, Gridded Population of the World (Balk et al. 2006), are constructed.

Tests and replications

In summary, each population distribution dataset produced in this study is characterized by input land cover data (AVHRR, MODIS, GLC2000 or GlobCover), a TCM (ADMIN-5, ADMIN-4, ADMIN-3, ADMIN-2, ADMIN-1, ADMIN-0, or no correction) and an EXL (RD, L5, L4, L3, L2 or L1) (Fig. 1).

In a first step, we fixed the extrapolation level to the maximum (i.e. EXL = L1)–because a high extrapolation level is likely to be required to produce population distribution data in other African countries–and varied the TCM. This allowed for exploration of the effectiveness of the population modelling procedures in the absence of high resolution census data. With EXL = L1, the sampling method is based on the Kenyan provinces. The selection of 4 out of the 8 provinces was replicated 25 times and these 25 different combinations (out of 70) were used to produce population distribution datasets. This was also repeated with the four land cover datasets as input data.

In a second step, we fixed the TCM and produced population distribution data for each of the 6 EXL. As sampling methods associated with EXL include a random component, each stage was replicated 25 times. This was repeated with the four land cover datasets as input data. In this second step, the TCM was fixed to ADMIN-2. To decide which level to use, we looked at the average spatial resolution of available census data in other African countries. On average, the census data available that is georegistered to administrative boundaries for African countries have an ASR of 84.88 km, which is closer to the district level (ADMIN-2) in Kenya (Table 3). Figure 2 shows the ASR of African countries.

Statistical analyses including analyses of variance and Tukey's honest significant difference tests were performed to test for differences between different land cover data, TCM and EXL. The Tukey's honest significant difference statistical test is used to identify which means are significantly different from the others. This test is based on the range of the sample means rather than the individual differences.

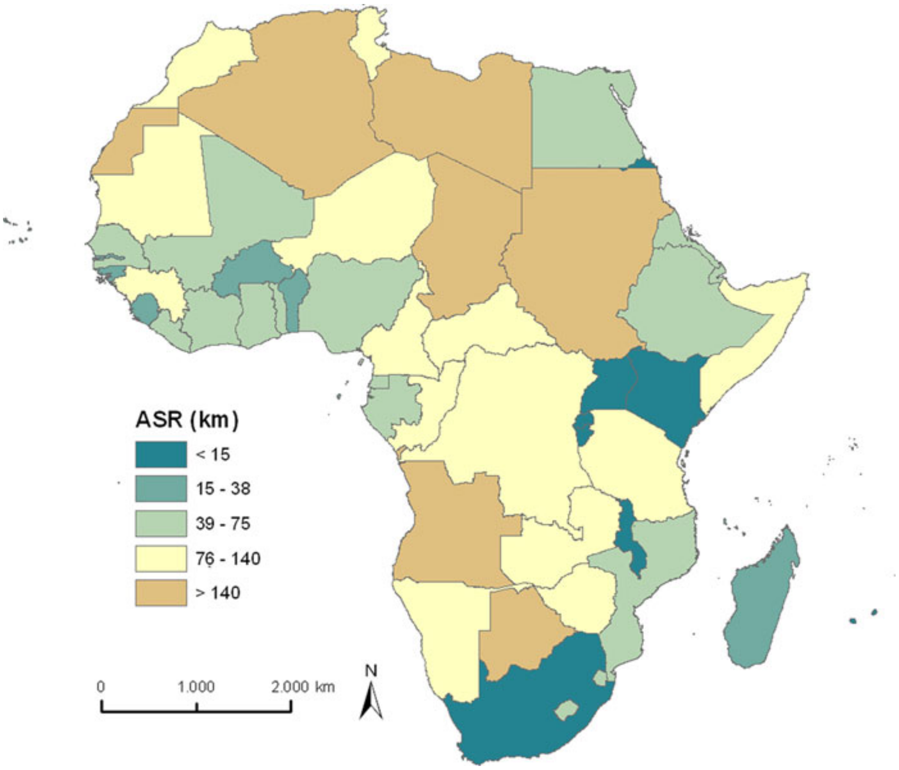

Fig. 2 Average spatial resolution (ASR) of census data used in the construction of Gridded Population of the World v3 (GPWv3) and the Global Rural Urban Mapping Project (GRUMP) in African countries. The ASR measures the effective resolution of administrative units in kilometers. It is calculated as the square root of the land area divided by the number of administrative units (Balk and Yetman 2004)

Results

Results from the first series of replications (with EXL fixed to L1) are presented in Figs. 3, 4. Firstly, Fig. 3 shows that in most of the cases (with TCM = ADMIN-4, ADMIN-3, ADMIN-2 and ADMIN-1), the GlobCover dataset used as input land cover data produced the lowest RMSE on average. An analysis of variance including the global land cover and the TCM as independent variables confirmed that the land cover dataset used in combination with detailed settlement extents had a significant impact on the RMSE (*F* value = 3.11; $p = 0.026$). Complete results from the analysis of variance are presented in table 4. The Tukey's test confirmed the significant difference between the GlobCover-based population distribution data and the AVHRR-based population data ($p = 0.016$). When removing the TCM = ADMIN-0 particular case, the Tukey's test showed significant differences between the GlobCover-based population data and all three other groups of population distribution data (with all *p*-values < 0.0001). The significant interaction factor shows that the effect of the choice of land cover is different according to the TCM level (Table 4).

The accuracy of population distribution data decreased drastically with coarser administrative levels used for TCM, both in terms of RMSE and correlation coefficient (Fig. 4). This is even more marked for ASR below 100 km. Without any correction by totals, the population distribution data produced show similar Pearson correlation coefficients as those of population data produced with TCM = ADMIN-0, but RMSE approximately 100 times higher, with average RMSEs between 411,467 for the GLC2000-based population model and 615,896 for the GlobCover-based population model.

Figure 3 also allows comparison of the accuracy of population distribution data produced with the areal weighting method (dotted line in the graphs). We observe that with TCM = ADMIN-5, the areal weighted method produced more accurate population

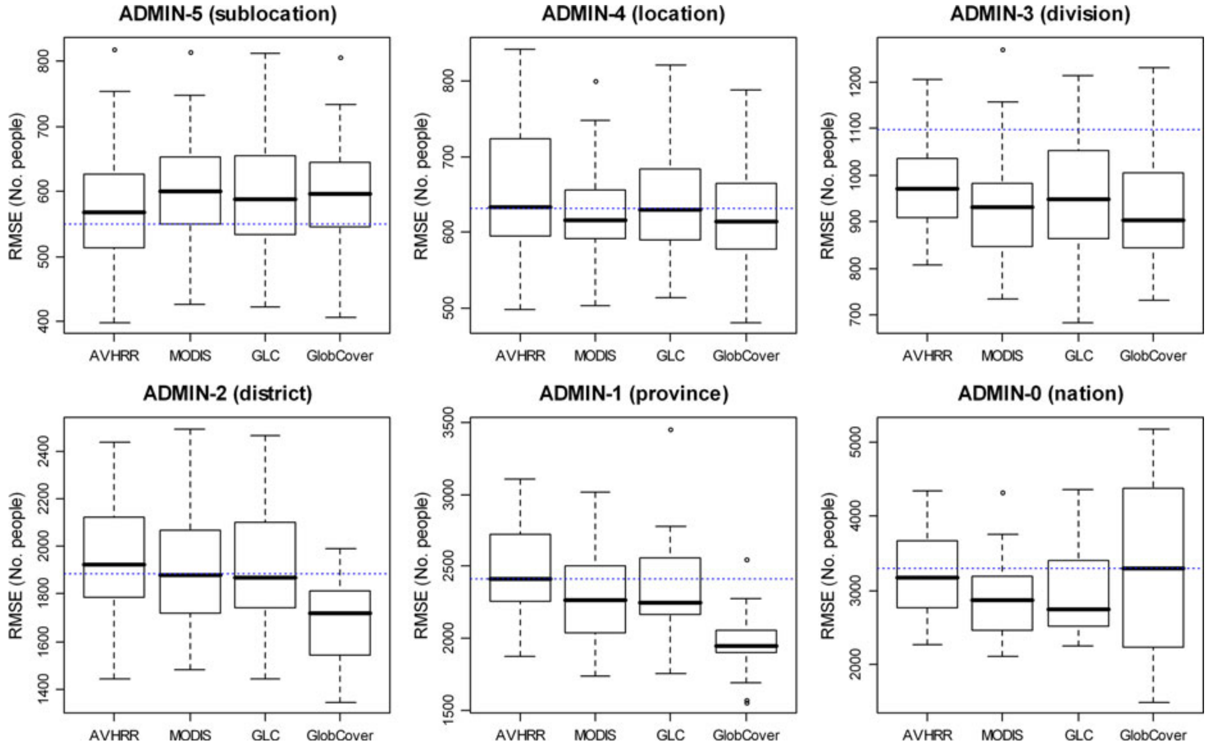

Fig. 3 Results from accuracy assessments of population distribution data produced with EXL = L1. *Boxplots* show the RMSEs according to the TCM and the global land cover data used as input data. Each stage was replicated 25 times. The *dotted line* corresponds to the RMSE associated with the areal weighted method (i.e. homogenous distribution of people within administrative units) for each administrative level

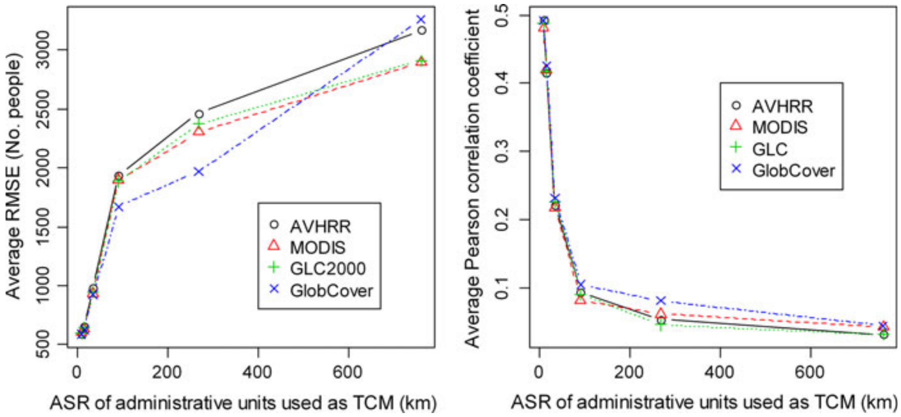

Fig. 4 Average RMSE and Pearson correlation coefficients as a function of the ASR of the 6 administrative levels in Kenya used for TCM in the population distribution modelling procedure

Table 4 Results from the analysis of variance performed on RMSEs extracted from population maps

		F value	*Pr* (> *F*)
a.	TCM	762.44	<0.0001
	LC	3.11	0.02586
	TCM:LC	2.98	0.00013
b.	TCM	1471.64	<0.0001
	LC	16.76	<0.0001
	TCM:LC	5.54	<0.0001

The analysis includes the TCM, the land cover dataset used in combination with detailed settlement extents for population distribution modelling (LC) and the interaction between these two factors (TCM:LC) as independent factors. EXL is here fixed to L1. The first analysis of variance (a) includes all TCM levels, whereas in the second analysis (b) the particular case TCM = ADMIN-0 was removed

distribution data than the procedure described in this paper, whereas for the other levels of TCM, the land cover based population data were generally more accurate. Moreover, for ADMIN-2 and ADMIN-1 levels of TCM, the improvement shown for the GlobCover-based population distribution dataset compared to the areal weighted data is much clearer than the population data based on other land cover datasets.

Results from the second series of replications (with TCM fixed to ADMIN-2) are presented in Fig. 5. This figure shows that the accuracy of population distribution models decreases slightly with an increasing level of extrapolation. Gridded data produced by extrapolating population densities from one province to the other (EXL = L1) provided the highest RMSEs on average. However, the differences between extrapolation levels is not significant at the 95% confidence level according to our analysis of variance, whereas the global land cover data used in the modelling procedure is still highly significant (Table 5). According to the Tukey's test, the GlobCover-based population distribution data are again significantly different from the population distribution data based on other global land cover data (with all p-values < 0.0001).

We performed 25 random simulations for each combination of TCM and EXL. Results showed that the average RMSE converged appreciably after this reasonable number of simulations, with changes in the average RMSE in the last 5 simulations generally lower than 2% and lower than 1% in 84% of cases.

Discussion

The primary aim of this work was to identify which global land cover data could be used in combination with detailed settlement extents to produce the most accurate population distribution modelling across Africa. Results showed that, combined with detailed settlement extents, the GlobCover dataset generally provided significantly more accurate population distribution models than other global land cover datasets in Kenya. As a massive majority of people across the World reside in settlements, it was important to refine global land cover datasets with as detailed as possible settlement extents data. However, we showed that different refined land cover data resulted in significantly different output population distribution datasets, which confirms that the use of additional land cover classes for dasymetric modelling can further improve population distribution models.

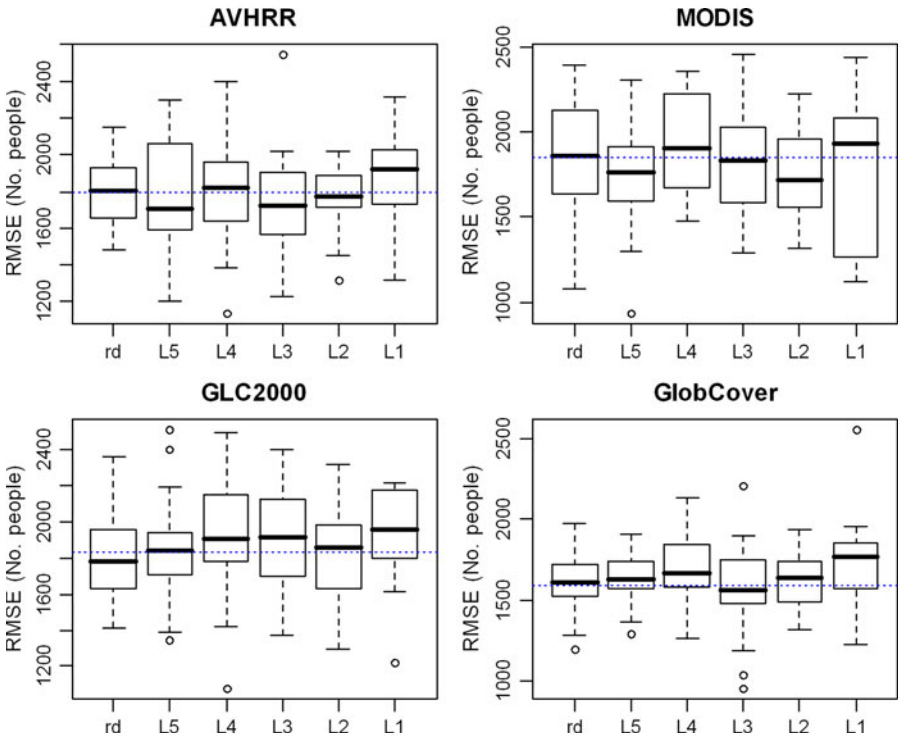

Fig. 5 Results from accuracy assessments of population distribution data produced with TCM = ADMIN-2. *Boxplots* show the RMSEs according to the EXL and the global land cover data used as input data. Each stage was replicated 25 times. The *dotted line* corresponds to the RMSE of the population data calculated without sampling method (i.e. all EAs were used for both modeling and accuracy assessment)

Table 5 Results from the analysis of variance performed on RMSEs extracted from population maps

	F value	*Pr* (> *F*)
EXL	2.00	0.07635
LC	17.37	<0.0001
EXL:LC	0.93	0.52857

The analysis includes the EXL, the land cover dataset used in population mapping (LC) and the interaction between these two factors (EXL:LC) as independent factors. TCM is here fixed to ADMIN-2

Because our ultimate objective is to extend the population modelling method to other African countries, we tested the effectiveness of the population modelling procedures in the absence of high resolution census data. In the majority of African countries, the most detailed census data available have an ASR higher than 75 km (Fig. 2). For such coarse resolution data (close to ADMIN-2 in Kenya, see Table 3), GlobCover clearly provided the most accurate results (Fig. 3, 4). Figure 6 shows an example of gridded population distribution data produced for Kenya based on the GlobCover dataset with TCM = ADMIN-2 and EXL = L1.

We also tested the effectiveness of the population distribution modelling procedures with different extrapolation levels. Our results showed no significant influence of the extrapolation level used on the accuracy of the output population distribution datasets for Kenya. This does not exclude accuracy losses when land cover specific population densities are extrapolated from one country to the other for large area population distribution modelling, as the relationship between population density and land cover differs from one country to the other. The spatial extrapolation level should therefore be minimized as much as possible in any large area population distribution modelling. Even if the impact of a high extrapolation level was limited in our analysis, whichever global land cover data used, population weights can only be extrapolated to spatially proximate and environmentally similar regions.

The better performance of the GlobCover dataset for population distribution modelling is most likely due to its finer spatial resolution (300 m compared to 1 km for AVHRR, MODIS and GLC2000). The GlobCover dataset also includes a larger number of land cover classes compared to other global land

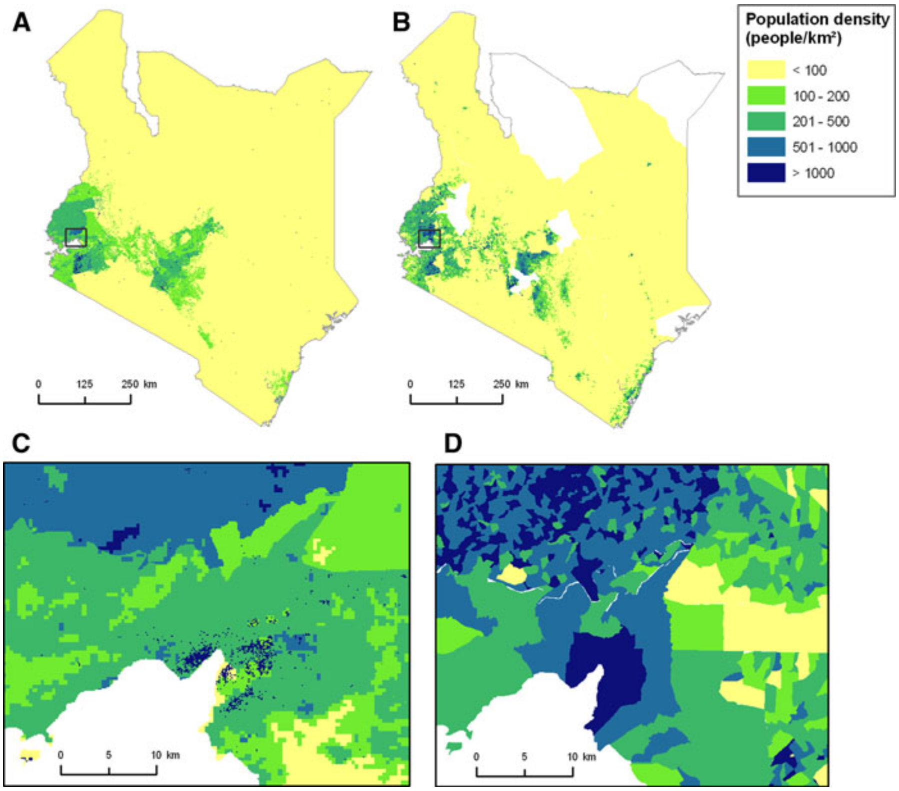

Fig. 6 Population distribution maps for Kenya. **a** example of population distribution predicted using GlobCover data for dasymetric modelling, with TCM = ADMIN-2 and EXL = L1. **b** enumeration area census data (observed data). **c** close-up of the population distribution map for Kisumu district. **d** close-up of enumeration area census data for Kisumu district

cover datasets (Table 1), which could enable greater precision in the derivation and modelling of land cover-population density relationships. However, a large number of different land cover classes would only improve the accuracy of population distribution data produced if population densities are significantly different by land cover class. The optimal land cover data for population distribution modelling would be a land cover classification that maximizes within land cover class homogeneity and maximizes between-class heterogeneity in relation to population density. In addition, the per land cover specific weights calculated can actually be less accurate with a higher number of classes. Combining land cover classes could therefore increase the accuracy of population distribution data. An additional analysis showed that combining GlobCover land cover classes did not however influence significantly the accuracies of output population distribution datasets here (see supplementary material).

GlobCover provided less accurate average results in the worst modelling situation, i.e. with the highest level of extrapolation (EXL = L1) and the lowest administrative level for the correction by the totals (TCM = ADMIN-0) (see the last boxplot in Fig. 3) or without any correction by the totals. In some particular cases, the population distribution data produced using the GlobCover dataset provided very high RMSEs, which increased the RMSE variation and reduced considerably the average accuracy. The large number of land cover classes in the GlobCover dataset made the per land cover class specific densities sometimes less accurate because they were calculated based on a limited number of EAs where these land cover classes are dominant. The RMSEs of GlobCover-based population data are thus higher in some particular situations, but the correlation coefficient is always higher on average for GlobCover-based population distribution data (Fig. 4). Aggregating land cover classes could limit this effect.

The time of land cover data acquisition may also influence the results. The date of imagery acquisition for MODIS and GLC2000 (2001 and 2000 respectively, see Table 1) were close to the census data (1999), whereas the AVHRR data were older (1981–1994) and the MERIS imagery used for GlobCover were more recent (2005–2006). With substantial population growth and urbanization taking place across Africa, the expansion of cities may have been important. In our analysis, the urban classes of land cover data have been refined based on the settlement map from Tatem et al. (2007), which relied upon data collected between 1999 and 2002. The discrepancy between census and land cover data is therefore limited for urban areas in our analysis. However, other land use changes, such as the expansion of cropland over natural vegetation may have changed in Kenya and could have induced discrepancies between census and land cover data.

This study supports prior work on population distribution modelling. Firstly, the gridded population distribution datasets produced from census data and satellite imagery derived land cover data generally provided more accurate results than areal weighting, as already shown in Tatem et al. (2007) and Mennis and Hultgren (2006). However, when using sublocation level census data in Kenya (with ASR < 10 km), the areal weighting method provided the most accurate results (first boxplot in Fig. 3). This suggests that when very fine-resolution census data are available, the use of land cover data at the spatial resolutions considered here in population distribution modelling does not necessarily improve the simple areal weighting method. This demonstrates that the approach only increases population distribution model accuracies over the simple gridding of census data if land cover data are significantly more detailed than the input census data. In our case, more spatially detailed ancillary data would be needed to improve the redistribution of populations within sublocation units in Kenya. Secondly, Fig. 4 shows the accuracy changes experienced with different ASR of census data available for modelling. As already described in Hay et al. (2005), it demonstrates that obtaining as high a spatial resolution of census data as possible must be the priority starting point in population distribution modelling. Given the resolution of census data available, the ancillary data can improve population model accuracies to a lesser extent. The potential improvement provided by land cover data is higher with coarser ASR of input census data.

In conclusion, GlobCover, in combination with detailed settlement extents, likely represents a more accurate source of land cover data for dasymetric modelling than other global land cover datasets. In addition, GlobCover is the most recent global land cover dataset, being derived from 2005/2006 MERIS imagery. Moreover, the robust automated processes used in the data production (Arino et al. 2007, 2008) allows for updates to be incorporated in the coming years. A complete land cover dataset for the year 2009 is currently under production (ESA GlobCover Team 2009). For all these reasons, GlobCover represents the preferred global land cover dataset for use as an alternative to regional land cover products in the creation of population distribution data across large areas.

These analyses form part of a wider initiative, the AfriPop project (www.afripop.org), aimed at providing detailed and open access gridded population distribution data for all African countries. AfriPop aims to produce datasets based on freely available data and methods that can easily incorporate new data as it becomes available.

Acknowledgments CL is supported by a grant from the Fondation Philippe Wiener - Maurice Anspach. AJT is supported by a grant from the Bill and Melinda Gates Foundation (#49446). This work forms part of the output of the AfriPop Project (www.afripop.org), principally funded by the Fondation Philippe Wiener–Maurice Anspach, and the Malaria Atlas Project (MAP, www.map.ox.ac.uk), principally funded by the Wellcome Trust, U.K.

References

Arino, O., Bicheron, P., Achard, F., Latham, J., Witt, R., & Weber, J. L. (2008). GLOBCOVER: The most detailed portrait of Earth. *European Space Agency, 136*, 24–31.

Arino, O., Gross, D., Ranera, F., Leroy, M., Bicheron, P., Brockman, C., Defourny, P., et al. (2007). GlobCover: ESA service for global land cover from MERIS. In *IEEE International Geoscience and Remote Sensing Symposium 2007* (pp. 2412–2415).

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*, 119–156.

Balk, D., Gorokhovich, Y., & Levy, M. (2005a). *Estimation of coastal populations exposed to the 26 december 2004 Tsunami*. New York: Working Paper, Center for International Earth Science Information Network, Columbia Univ.

Balk, D., Pozzi, F., Yetman, G., Deichmann, U., & Nelson, A. 2005b. The distribution of people and the dimension of place: Methodologies to improve the global estimation of urban extents. In International Society for Photogrammetry and Remote Sensing, Proceedings of the Urban Remote Sensing Conference.

Balk, D., Storeygard, A., Levy, M., Gaskell, J., Sharma, M., & Flor, R. (2005b). Child hunger in the developing world: An analysis of environmental and social correlates. *Food Policy, 30*(5–6), 584–611.

Balk, D., & Yetman, G. (2004). The Global Distribution of Population: Evaluating the gains in resolution refinement. Center for International Earth Science Information Network (CIESIN). http://sedac.ciesin.org/gpw/docs/gpw3_documentation_final.pdf. Accessed 2 February 2010.

Bhaduri, B., Bright, E., Coleman, P., & Dobson, J. (2002). LandScan: Locating people is what matters. *Geo Informatics, 5*(2), 34–37.

Bhaduri, B., Bright, E., Coleman, P., & Urban, M. L. (2007). LandScan USA: A high-resolution geospatial and temporal modeling approach for population distribution and dynamics. *GeoJournal, 69*(1), 103–117.

CBS. (2001). *1999 Population and housing census: Counting our people for development. Volume 1: Population distribution by administrative areas and urban centres. Central Bureau of Statistics (CBS), ministry of finance and planning*. Nairobi, Kenya: Government of Kenya.

Deichmann, U. (1996). *A review of spatial population database design and modeling*. Santa Barbara, California, USA: National Center for Geographic Information and Analysis (NCGIA), University of California, Santa Barbara (UCSB).

Deichmann, U., Balk, D., & Yetman, G. (2001). *Transforming population data for interdisciplinary usages: From census to grid*. Washington (DC): Center for International Earth Science Information Network.

Dobson, J. E., Bright, E. A., Coleman, P. R., Durfee, R. C., & Worley, B. A. (2000). LandScan: A global population database for estimating populations at risk. *Photogrammetric Engineering and Remote Sensing, 66*(7), 849–857.

ESA GlobCover Team (2009). GlobCover newsletter no9, october.

Ferguson, N. M., Cummings, D. A., Cauchemez, S., Fraser, C., Riley, S., Meeyai, A., et al. (2005). Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature, 437*(7056), 209–214.

Friedl, M. A., McIver, D. K., Hodges, J. C. F., Zhang, X. Y., Muchoney, D., Strahler, A. H., et al. (2002). Global land cover mapping from MODIS: Algorithms and early results. *Remote Sensing of Environment, 83*(1), 287–302.

Fritz, S., Bartholome, E., Belward, A., Hartley, A., Stibig, H. J., Eva, H., et al. (2002). *Harmonisation, mosaicing and production of the Global Land Cover 2000 database (Beta Version)*. Brussels (Belgium): Directorate-General, Joint Research Centre, European Commission.

Guerra, C. A., Snow, R. W., & Hay, S. I. (2006). Defining the global spatial limits of malaria transmission in 2005. *Advances in Parasitology, 62*, 157–179.

Hansen, M. C., DeFries, R. S., Townshend, J. R. G., & Sohlberg, R. (2000). Global land cover classification at 1 km spatial resolution using a classification tree approach. *International Journal of Remote Sensing, 21*(6), 1331–1364.

Hay, S.I., Guerra, C.A., Gething, P.W., Patil, A.P., Tatem, A.J., Noor, A.M., Kabaria, C.W., Manh, B.H., Elyazar, I.R.F., Brooker, S., Smith, D.L., Moyeed, R.A., & Snow, R.W. (2009). A world malaria map: Plasmodium falciparum endemicity in 2007. *Public Library of Science Medicine*, 6(3): e48.

Hay, S. I., Noor, A. M., Nelson, A., & Tatem, A. J. (2005). The accuracy of human population maps for public health application. *Tropical Medicine and International Health, 10*(10), 1073.

Hyman, G., Lema, G., Nelson, A., & Deichmann, U. (2004). *Latin American and Caribbean population database documentation*. Mexico: International Center for Tropical Agriculture.

Maynard-Ford, M. C., Phillips, E. C., & Chirico, P. G. (2008). Mapping vulnerability to disasters in Latin America and the Caribbean, 1900–2007.

McGranahan, G., Balk, D., & Anderson, B. (2007). The rising tide: Assessing the risks of climate change and human settlements in low elevation coastal zones. *Environment and Urbanization, 19*(1), 17.

Mennis, J. (2003). Generating surface models of population using dasymetric mapping. *The Professional Geographer, 55*(1), 31–42.

Mennis, J., & Hultgren, T. (2006). Intelligent dasymetric mapping and its application to areal interpolation. *Cartography and Geographic Information Science, 33*(3), 179–194.

Nelson, A. (2004). African population database documentation. http://na.unep.net/globalpop/africa/. Accessed 2 February 2010.

Nicholls, R. J., Tol, R. S. J., & Vafeidis, A. T. (2005). Global estimates of the impact of a collapse of the West Antarctic Ice Sheet. *Climatic Change, 91*(1–2), 171–191.

Rao, D. M., Chernyakhovsky, A., & Rao, V. (2009). Modeling and analysis of global epidemiology of avian influenza. *Environmental Modelling and Software, 24*(1), 124–134.

Rogers, D. J., Wilson, A. J., Hay, S. I., & Graham, A. J. (2006). The global distribution of yellow fever and dengue. In S. I. Hay, A. Graham & D. J. Rogers (Eds.), *Global mapping of infectious diseases: methods, examples and emerging applications. Advances in Parasitology* (Vol. 62, pp. 181–220). London, UK: Academic Press.

Salvatore, M., Pozzi, F., Ataman, E., Huddleston, B., & Bloise, M. (2005). *Mapping global urban and rural population distributions*. Rome: FAO.

Storeygard, A., Balk, D., Levy, M., & Deane, G. (2008). The global distribution of infant mortality: A subnational spatial view. *Population Space and Place, 14*(3), 209–229.

Tatem, A. J., Guerra, C. A., Kabaria, C. W., Noor, A. M., & Hay, S. I. (2008). Human population, urban settlement patterns and their impact on Plasmodium falciparum malaria endemicity. *Malaria Journal, 7*(1), 218.

Tatem, A. J., Noor, A. M., & Hay, S. I. (2004). Defining approaches to settlement mapping for public health management in Kenya using medium spatial resolution satellite imagery. *Remote Sensing of Environment, 93*(1–2), 42–52.

Tatem, A. J., Noor, A. M., & Hay, S. I. (2005). Assessing the accuracy of satellite derived global and national urban maps in Kenya. *Remote sensing of environment, 96*(1), 87–97.

Tatem, A. J., Noor, A. M., von Hagen, C., Di Gregorio, A., & Hay, S. I. (2007). High resolution population maps for low income nations: Combining land cover and census in East Africa. *PLoS One*, *2*(12). doi:10.1371/journal.pone.0001298.

Tobler, W., Deichmann, U., Gottsegen, J., & Maloy, K. 1995. The global demography project. *Department of geography, university of California, Santa Barbara, NCGIA technical report TR-95-6*.

Tobler, W., Deichmann, U., Gottsegen, J., & Maloy, K. (1997). World population in a grid of spherical quadrilaterals. *International Journal of Population Geography, 3*(3), 203–225.

Wright, J. K. (1936). A method of mapping densities of population: With Cape Cod as an example. *Geographical Review, 26*, 103–110.