UN/POP/EGM-CPD49/2015/15

ENGLISH ONLY

UNITED NATIONS EXPERT GROUP MEETING ON STRENGTHENING THE DEMOGRAPHIC EVIDENCE BASE FOR THE POST-2015 DEVELOPMENT AGENDA
**Population Division**
**Department of Economic and Social Affairs**
**United Nations Secretariat**
**New York**
**5-6 October 2015**

# MAPPING POPULATION NUMBERS, DEMOGRAPHICS AND BEHAVIOURS: THE WORLDPOP PERSPECTIVE[1]

*Andrew J Tatem*[*]

[1]**The opinions expressed in this paper are those of the author and do not necessarily reflect those of the United Nations or its Member States. This paper is reproduced as submitted by the author. This paper is being reproduced without formal editing.**

[*] **University of Southampton/WorldPop *(*www.worldpop.org).**

# Mapping population numbers, demographics and behaviours: The WorldPop perspective

*Prof. Andrew J Tatem, University of Southampton/WorldPop (www.worldpop.org)*

Measuring progress towards international health and development goals requires a reliable baseline from which to measure change, and recent increases in spatially referenced data and methodological advancements have advanced our abilities to measure, model and map the presence and prevalence of many key indicators using sophisticated spatial tools. The provision of burden or population at risk estimates generally requires linking these estimates with spatial demographic data, but for many resource-poor countries, contemporary and regularly updated subnational data on total population sizes, distributions, compositions and temporal trends are lacking, prompting a reliance on uncertain estimates. WorldPop (www.worldpop.org), has worked with ministries of health, statistics agencies and other organizations over the past decade to attempt to fill these gaps, drawing on traditional and novel data sources to produce high spatial resolution open-access demographic data sets.

**Figure 1. Challenges of using multiple data sources**

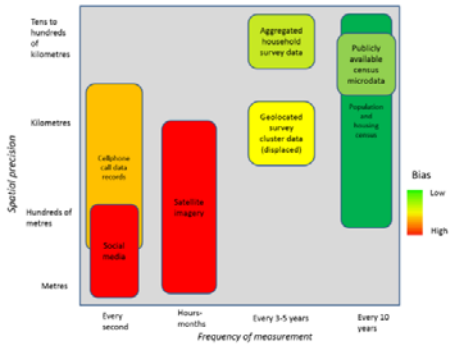



*Fig 1 – examples of datasources used in the construction of high resolution population maps in low/middle-income countries, and their features*

*Source*: Developed by author.

The basis for the vast majority of our current knowledge on the spatial distribution and composition of populations are censuses. National censuses can provide a comprehensive and relatively unbiased source of information at a single time point (figure.1), and when linked with accurate boundary data, provide a spatially detailed evidence base on population. Further processing of these data through integration with higher spatial resolution 'covariate' datasets in modelling frameworks, can then disaggregate these boundary-linked counts to consistent gridded representations (figure.2), Stevens and others, 2015; Balk and others, 2006; Azar and others, 2013; Bhaduri and others, 2002; Sorichetta and others, 2015). Moreover, while barriers and sensitivity issues remain in accessing the most granular levels of census data at enumeration area and individual levels, efforts are being made to improve access through microdata samples, representative at subnational levels (figure 1), https:// international.ipums.org/international, www.terrapop.org).

Conducting a national census is however an arduous, resource-intensive undertaking and is a challenge even in countries with the necessary technology, infrastructure and financial and human capacity. In low income nations, or those that have undergone internal strife, civil wars and frequent changes in government, up-to-date and maintaining accurate census data is an extraordinary challenge. Even in countries where a quality census has occurred in the past, keeping up with rapid and inconsistent population growth remains difficult. As a result, in many low income nations, existing census data is generally outdated, leaving both the host governments and the global community without a reliable source of population denominators at subnational scales. The impact of this "data deficiency" was painfully apparent with the recent Ebola outbreak in West Africa, where emergency

responders struggled to identify the location and size of rural settlements and could not accurately calculate infection rates since the denominator was not known, an issue that is regularly encountered (Hillson and others, 2014; Tatem, 2014). Nearly all public health outreach efforts, for example, from vaccinations to bed nets to HIV treatment, depend on accurate target population denominators to not only estimate project costs and resource needs, but also to measure and assess results and impact. The MDGs and upcoming SDGs are all based on ensuring a certain percentage of the population has access to specific services or resources, or achieves a certain level of social, economic, or physical health. These measurements require a solid and regularly updated understanding of not only how many people live in a country, but where the people are, and who they are.

**Figure 2. WorldPop population mapping example**

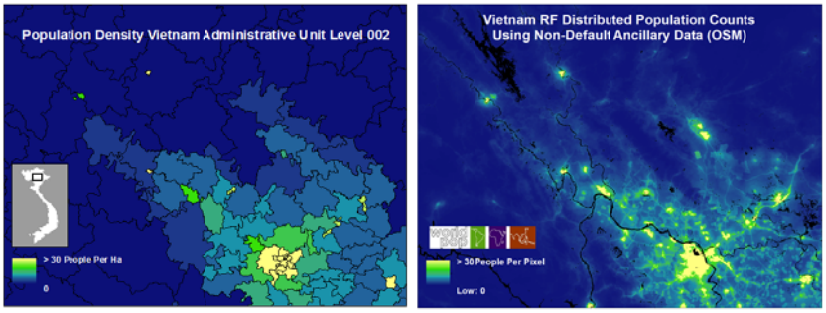
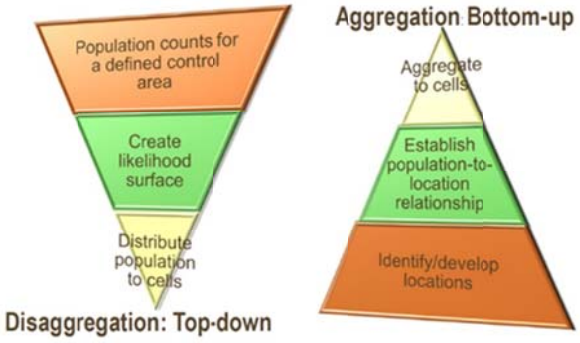



*Fig.2. WorldPop population mapping example: (left) Population density from census data for each administrative level 2 unit in an area of northern Vietnam, (right) WorldPop population modelling methods take the census data as input, then use machine learning methods to exploit the relationship between population density and high resolution landscape features, such as those from land cover and satellite data, to predict population densities for each 100x100m grid cell on the landscape.*

*Source*: WorldPop (2015). Available from www.worldpop.org.

**Figure 3. Conceptual approaches to producing gridded population maps**




*Source*: Seaman, Vincent (2015) of Bill and Melinda Gates Foundation. Personal Communication.

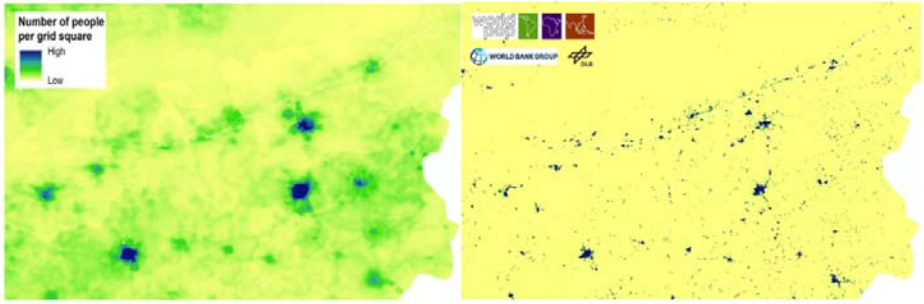
Such requirements and the deficiencies of national census data mean that other data sources are increasingly being explored in efforts to produce estimates at different geographical scales and time periods. Figure 1 highlights some of those being utilised within WorldPop to complement census data in the detailed mapping of populations and their characteristics across timescales, and for which examples are provided in the remainder of this document. Though increasingly prone to bias through measurement of smaller sample sizes (for example, geo-located household survey clusters), specific demographic groups (for example, social media) or simply factors related to population densities (for example, satellite imagery), each source has advantages over census data in terms of the frequency of measurement and spatial precision (figure 1). Moreover, their utilization represents a gradual shift from “top-down” approaches where census data counts are maintained and disaggregated to small

areas, to more “bottom-up” approaches, where estimates are made independent of census data (figure 3).

## A. Spatial disaggregation: Top-down approaches

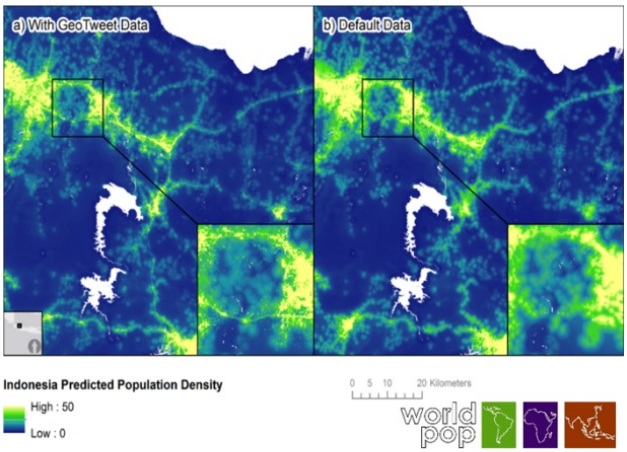
Where contemporary and reliable census data matched to accurate, detailed boundaries exist, top-down approaches (figure 3) to population count and composition mapping are valuable in providing an unbiased and precise picture of populations. Spatial covariate datasets used in the population disaggregation process tend to include factors known to correlate with population densities, such as satellite-derived maps of human settlements, urban areas, topography, lights at night, and land cover. Additionally, infrastructure-related variables have been used, including road networks and health facilities (for example, Stevens and others, 2015; Sorichetta and others 2015, figure 2). Increasingly, high resolution satellite imagery processed using sophisticated image analysis techniques, are enabling the large scale mapping of built up areas and individual buildings at fine spatial detail (Esch and others, 2011; Pesaresi and others, 2011). These unprecedented levels of detail in global human settlement mapping are resulting in knock-on effects in the spatial detail and accuracy of top-down population mapping methods (for example, figure 4), and will form the basis for WorldPop global population mapping over the coming months in collaboration with the Bill and Melinda Gates Foundation, Joint Research Center of the European Commission, World Bank and Center for International Earth Science Information Network.

**Figure 4. Predicted population map for an area in Guatemala using WorldPop mapping methods (left) without inclusion of global urban footprint and (right) including global urban footprint as a covariate**




Source: www.worldpop.org.

**Figure 5. Improvements in population mapping in Indonesia through using geotweet densities as a covariate**




*Source*: Patel, N. and others (2015).

Nevertheless, all of these covariates are typically static in nature and are not direct measures of the presence of people. The rise in data availability of user communications and check-ins through social media, such as Twitter, presents opportunities however, in terms of a data source that is freely available and dynamic. Although highly biased towards certain demographic groups (figure 1), and on an average day only about 1.6 per cent of tweets are posted with an exact geolocation, many useful geographic applications have been derived from tweet data (Leetaru and others, 2013; Hawelka and others, 2014). The maps of geo-located tweets in countries where Twitter is popular show detailed depictions of human activity, with the location of tweets indicative of settlements, transportation networks, and building locations (Leetaru and others, 2013; figure 5). The integration of these data as a covariate into approaches for the disaggregation of admin-unit based population counts shows great potential in terms of improvements in mapping accuracies (Patel and others, 2015; figure 5). As smartphones continue to proliferate, the results underline the potential of this data source in contributing to the improvement of population mapping and its dynamic update (Leetaru and others, 2013). Furthermore, other sources of social media data, some country specific like Baidu (China), Instagram, Shutterfly, and others, also offer potential when the data is not only geospatially referenced but made freely available for research.

## B. BOTTOM-UP APPROACHES

Modern technology is offering solutions to tackling the gaps in our knowledge of population numbers and distributions in the resource poor regions where census data are unreliable, outdated or of coarse resolution. The mapping of human settlements and even individual buildings from new generations of satellites or from aerial photography is providing detailed geospatial data on the human landscape (Hillson and others, 2014; Graesser and others, 2012). Computer vision and machine learning approaches can then distinguish typologies of patterns settlement (for example, Graesser and others, 2012), and when combined with estimates of occupancy from ground surveys, these offer a 'bottom-up' approach to population size estimation and mapping that potentially circumvents the requirement for census data (for example, Hillson and others, 2014). Such an approach is being tried in Nigeria to support vaccination planning and resource allocation efforts (figure 6, http://vts.eocng.org) and soon to be adapted to other countries.

**Figure 6. Population mapping using remote sensing and ground surveys**

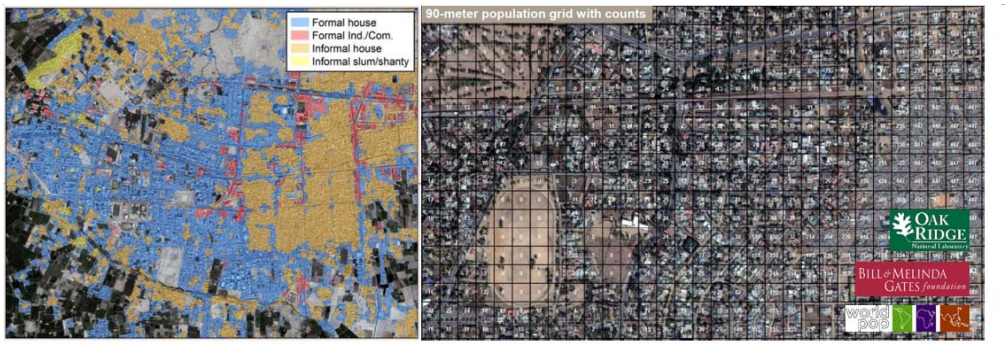


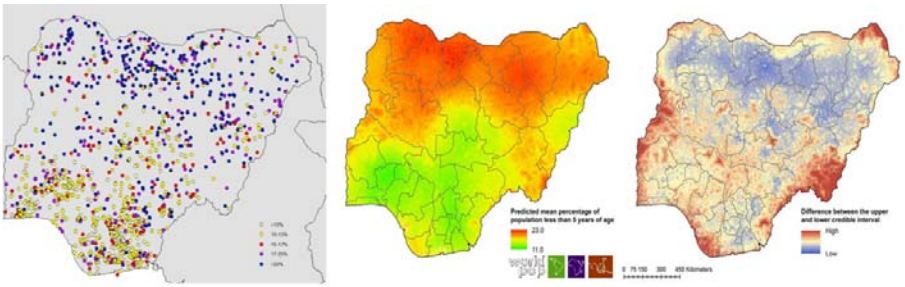

*Fig.6. (left) high resolution settlement extraction and typology derived from satellite imagery; (right) population estimate per grid cell through linking the satellite-derived typologies with ground surveys. Data used in http://vts.eocng.org.*

*Source*: Seaman, Vincent (2015) of Bill and Melinda Gates Foundation. Personal Communication.

Vulnerable groups such as under-five-age group, women of childbearing age and the elderly remain the focus of the Millennium Development Goals (MDGs) and Sustainable Development Goals (SDGs). Previous approaches to estimating vulnerable populations at risk in influential health studies have been limited by data availability and simply taken existing spatial population count data and applied national level multipliers (for example, Murray and others, 2012; Garske and others, 2014). Analyses have shown that, on top of the existing issues with total population counts and distributions

built on census data in resource poor settings, such an approach leads to significant differences in vulnerable population at risk estimates over accounting for the subnational variations that are universal in population age structures (Tatem and others, 2013). The growth in national household surveys however, including the availability of cluster-level GPS coordinates (for example, figure 7) are providing new contemporary and more spatially detailed data for improving estimates of vulnerable population distributions.

**Figure 7. WorldPop model-based high resolution geostatistical population mapping**



*Flg.7.WorldPop high resolution mapping of population age structures in the absence of census data: (left) geolocated household survey cluster data coloured by the proportion of the population surveyed that was under 5 years of age; (middle) predicted proportion of the population under 5 years of age per 1x1km grid square using model-based geostatistics in a Bayesian framework; (right) map of per grid square uncertainty measure showing the level of confidence in each prediction made in the under 5yr proportion map. Adapted from Alegana et al (2015).*

*Source:* Alegana, V. and others (2015).

Increased availability and use of geolocated cluster survey data has also coincided with greater recognition by policymakers and researchers of the need for valid approaches to estimating health and population indicators in small administrative areas such as districts, counties and other sub-provincial areas—areas smaller than the usual regions upon which the sampling is powered to represent (DHS Spatial Interpolation Working Group, 2014). Two approaches currently exist that allow collection of indicator estimates for these small administrative areas. The first is scaling up the data collection process by increasing the sample size, survey costs, and survey time to create a representative sample at the desired administrative level. The second is spatial interpolation using modelling techniques to predict values at non-surveyed locations. Given that the first approach is not feasible in an increasingly resource-constrained environment, the second approach, which uses spatial modelling techniques, is growing in popularity and has taken precedence in programmes such as the Demographic and Health Surveys (DHS Spatial Interpolation Working Group, 2014). Here, both the spatial relationships between variables measured at the cluster level and the relationships with spatially-detailed covariate layers (such as satellite-derived land use and settlement maps, and GIS data on infrastructure) are exploited in a Bayesian model-based geostatistical framework to map the variable of interest along with associated uncertainty metrics. Recent WorldPop collaborative projects have shown the potential of such approaches for mapping population age structures (figure 7; Alegana and others, 2015), poverty (figure 8; Bird and others, 2015), and literacy and sanitation (figure 8). Such approaches remain limited by data availability and the strength of the models, but the potential for mapping and monitoring progress towards development goals using them has recently been powerfully illustrated through pooling geolocated survey data on malaria prevalence (Bhatt and others, 2015).

Measuring change and providing reliable denominators across multiple time points and over large geographic areas represents a final challenge. A census or national household survey only records residential populations in a single snapshot (figure 1), without providing any detail of the daily, weekly and seasonal dynamics of population movements within countries. This means it's difficult to accurately assess the number of people who may be affected by, for example, climate change, or conflict, or for ascertaining in which direction a disease like Ebola is likely to spread. The proliferation of mobile phones (MPs) offers an unprecedented solution to this data gap. The global MP penetration rate (that is, the percentage of active MP subscriptions within the population) reached 96 per cent in 2014 (ITU, 2014). In developed countries, the number of MP subscribers surpasses the

total population, with a penetration rate now reaching 121 per cent, while in developing countries it is as high as 90 per cent, and continuing to rise (ITU, 2014). MP networks, also called cellular networks, are composed of cells, that is, geographic zones around a phone tower. Each MP communication can be located by identifying the geographic coordinates of its transmitting tower and the associated cell (Deville and others, 2014; Gonzales and others, 2008). This network-based positioning method is simple to implement and its accuracy directly depends upon the network structure, the higher the density of towers, the higher the precision of the MP communication geo-localization. Records detailing the time and associated cell of calls and text messages from de-identified users therefore provide a valuable indicator of human presence, and coupled with the increasing use of MPs, offer a promising alternative data source for increasing the spatial and temporal detail of large-scale population data sets.

**Figure 8. WorldPop Bayesian geostatistical 1x1km mapping of population characteristics from GPS-located household survey cluster data**

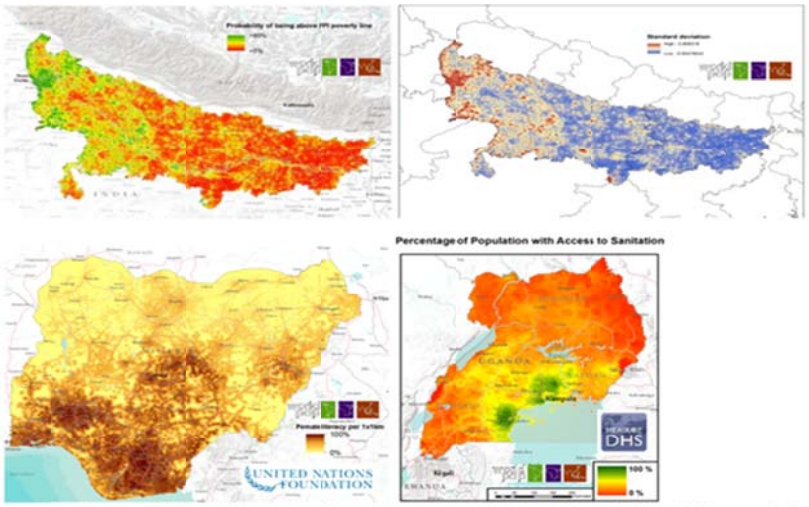



*Fig.8. Bayesian geostatistical 1x1km mapping of population characteristics from GPS-located household survey cluster data: (top-left) predicted mean poverty in northern India measured using the progress out of poverty index (PPI); (top-right) standard deviation in PPI estimates; (bottom-left) predicted mean female literacy rate in Nigeria; (bottom-right) predicted percentage of the population with access to sanitation in Uganda.*

*Source*: Gething, Peter and others (2015); Bird, T. and others (2015); and www.worldpop.org.

Recent analyses have shown that by using just a small segment of de-identified data that is already stored by mobile network operators, it is possible to rapidly produce detailed and up-to-date population distribution maps, sidestepping the need for the cumbersome once-a-decade census in providing such data (Deville and others, 2014; figure 9). Moreover, by using only phone call activity aggregated by tower, neither individual data nor connections between towers are used, guaranteeing the privacy of MP users, and overcoming concerns about data sensitivity. The analysis of MP data that is already collected every day by phone network providers can complement traditional census outputs, both in spatial and temporal terms. Not only can population maps of comparable accuracy to census data and existing downscaling methods be constructed solely from MP data (Deville and others, 2014), but these data offer additional benefits in terms of the measurements of population dynamics (for example, www.flowminder.org). Further, a combination of both the MP and remote sensing-based methods (Stevens and others, 2015) facilitates the improvement of both spatial and temporal resolutions and demonstrates how high resolution population datasets can be produced for any time period (Deville and others, 2015).

As these phone call records are continually collected, this also provides unprecedented insight into the nature of human population dynamics. Distribution maps can be easily drawn up for any period required—for example, day versus night, weekday versus weekend, workday versus holiday

difference. At little cost, this can help answer the type of questions that have previously been logistically challenging: how did population movements drive a cholera outbreak (Bengtsson and others, 2015; figure 10)? How have people reacted in the days following a devastating earthquake (www.worldpop.org.uk/nepal; figure 10)? How do health facility catchment population sizes change seasonally (Erbach-Schoenberg and others, 2015; Tatem and others, 2014)? A facilitated access to anonymized and aggregated forms of these data would greatly improve our knowledge of human population distribution and movements. Network providers are reticent to share their data because of privacy and marketing concerns. However, analyses have shown that aggregated and de-identified MP data could cost-effectively provide accurate maps of population distribution for every country in the world for every month (Deville and others, 2014). This processing could become a routine step, providing a substantially improved and timely understanding of population spatiotemporal dynamics, and will become a core activity of WorldPop-Flowminder over the coming year (www.worldpop.org.uk/about_our_work/case_studies).

**Figure 9. Mapping population density dynamics using cellphone data (circa 2014)**

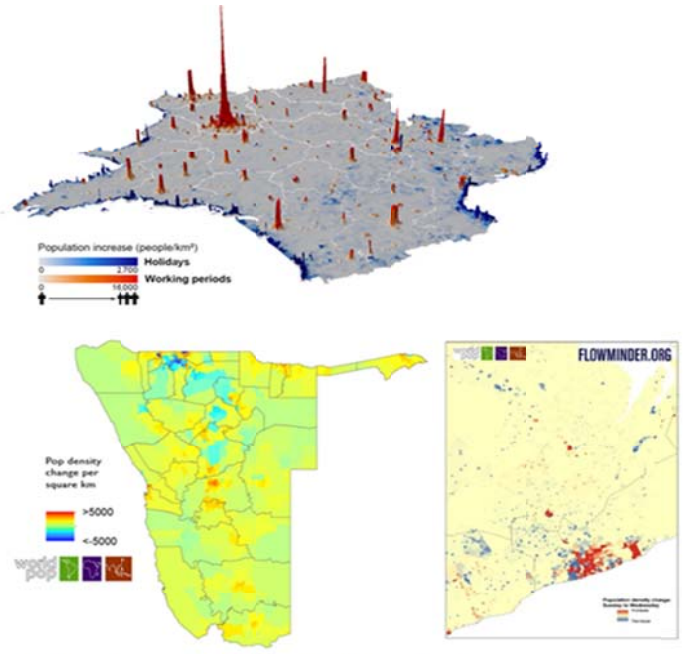
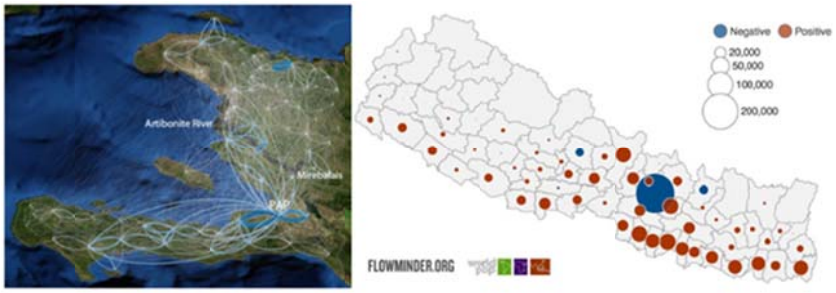



*Fig.9. Mapping population density dynamics using cellphone data:*
*(top) population density changes in France during holiday and work periods;*
*(bottom-left) population density change in early January in Namibia;*
*(bottom-right) population density change from weekend to weekday in the Accra region of Ghana.*

*Source*: Deville, P., and others (2014); Erbach-Schoenberg, E., and others (2015).

**Figure 10. Mapping population dynamics in (left) Haiti and (right) post-quake Nepal**




*Source*: Bengtsson, L., and others (2015). Available from www.worldpop.org.uk/nepal.

## 1. *Future directions and recommendations*

Spatial demographic data sets and production methods are rapidly improving, fuelled by improvements in technology and computing, but substantial limitations and uncertainties remain, particularly for those regions of the world where little data exists on how many people there are, where they live and their characteristics. Such uncertainties inherent in the demographic data sets used to provide denominators and processing steps taken are rarely acknowledged or accounted for, resulting in hidden uncertainties in many high impact development and health burden studies that are guiding international policies. In order to be able to measure progress in tracking development goals effectively, there is need for both methods to quantify the uncertainty inherent in spatial demographic data, and reliable denominator baselines from which to measure from—at present, for many of the resource-poor regions of the world these are still lacking, but as highlighted here, many opportunities exist. The integration of different forms of data can build on the strengths of each to overcome and account for weaknesses, gaps and biases (figure 1). Examples already exist, such as those outlined above that integrate cellphone, satellite, census and survey data, but the full potential has yet to be realised. The integration of these data into rigorous and robust spatiotemporal demographic modelling frameworks, with full quantification of uncertainty, represents an important next step, together with the strengthening of national statistical capacity to continue to provide high quality baseline data.

## RFERENCES


Alegana, V.A., and others (2015). Fine resolution mapping of population age-structures for health and development applications. *Journal of the Royal Society Interface*, vol. 12, No.105, (March 2015). Available from DOI: 10.1098/rsif.2015.0073.

Azar D., and others (2013). Generation of fine-scale population layers using multi-resolution satellite imagery and geospatial data. *Remote Sensing of Environment*, vol. 130, pp. 219-232.

Balk, D. L., and others (2006). Determining global population distribution: Methods, applications and data. *Advances in Parasitology*, vol. 62, pp. 119–156. Available from doi:10.1016/S0065-308X(05)62004-0.

Bengtsson, L., and others (2015). Using mobile phone data to predict the spatial spread of cholera. *Scientific Reports*, vol. 5, No. 8923.

Bhaduri, B.L., and others (2002). LandScan: Locating people is what matters. *Geoinformatics*, vol. 5, No. 2, pp. 34-37.

Bharti, N., and others (2015). Remotely measuring populations during a crisis by overlaying two data sources. *International Health*, vol. 7, No. 2, pp. 90-98.

Bhatt, S., and others (2015). The effect of malaria control on Plasmodium falciparum in Africa between 2000 and 2015. *Nature*, vol. 526, pp. 207–211. Available from doi:10.1038/ nature15535.

Bird, T., and others (forthcoming). High resolution progress out of poverty (PPI) mapping. *World Bank Economic Review*.

Deville, P., and others (2014). Dynamic population mapping using mobile phone data. *Proceedings of the National Academy of Sciences.* Available from doi:10.1073/pnas.1408439111.

DHS Spatial Interpolation Working Group (2014). Spatial interpolation with demographic and health survey data: Key considerations. DHS Spatial Analysis Reports No. 9. Rockville, Maryland, USA: ICF International.

Douglass, R.W., and others (2015). High resolution population estimates from telecommunications data. *EPJ Data Science*, vol. 4, No.4.

Ebener, S., and others (2015). The geography of maternal and newborn health: The state of the art. *International Journal of Health Geographics*, vol. 14, No. 19.

Esch, T., and others (2011). The path to mapping the global urban footprint using TanDEM-X data. *Proc ISPRS*, vol. 34.

Erbach-Schoenberg, E., and others (2015). Estimating variability in health facility catchment population sizes using mobile phone call records. Proceedings of NetMob 2015, Boston USA.

Garske, T., and others (2014). Yellow fever in Africa: Estimating the burden of disease and impact of mass vaccination from outbreak and serological data. *PLoS Medicine*, vol. 11, No. 5, e1001638.

Gething, Peter., and others (2015). Creating Spatial Interpolation Surfaces with DHS Data. DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International.

Gonzalez, M.C., C.A. Hidalgo and A.L. Barabasi (2008). Understanding individual human mobility patterns. *Nature*, vol. 453, pp. 779-782.

Graesser, J., and others (2012). Image based characterization of formal and informal neighbourhoods in an urban landscape. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 5, No. 4.

Hawelka, B., and others (2014). Geo-located Twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science*, vol. 41, No. 3, pp. 260-271. Available from doi:10.1080/15230406.2014.890072.

Hillson, J.D., and others (2014). Methods for determining the uncertainty of population estimates derived from satellite imagery and limited survey data: A case study of Bo City, Sierra Leone. *PLoS ONE*, vol. 9, No. 11. Available from doi:10.1371/journal.pone.0112241.

International Telecommunication Union (ITU) (2014) The World in 2014: ICT Facts and Figures. Available from http://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx. Accessed 5 May 2014.

Leetaru, K., and others (2013). Mapping the global Twitter heartbeat: The geography of Twitter. *First Monday*, vol. 18, No. 5.

Murray, C.J., and others (2012). Global malaria mortality between 1980 and 2010: A systematic analysis. *Lancet*, vol. 379, pp. 413–431.

Pesaresi, M., and others (2011). Toward global automatic built-up area recognition using optical VHR imagery. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, vol. 4, No.4, pp. 923-934.

Patel, N., and others (forthcoming). Improving large area population mapping using geotweet densities, *Transactions in GIS.*

Sankoh, O., and P. Byass (2012). The INDEPTH Network: filling vital gaps in global epidemiology. *The International Journal of Epidemiology*, vol. 41, No. 3, pp. 579-588.

Sorichetta, A., and others (2015). High-resolution gridded population datasets for Latin America and the Caribbean in 2010, 2015, and 2020. *Scientific Data*, vol.2, No. 150045 (2015). Available from doi:10.1038/sdata.2015.45.

Stevens, F.R., and others (2015). Disaggregating census data for population mapping using random forests with remotely-sensed and ancillary data. *PLoS ONE*, vol. 10, No. 2, e0107042. Available from doi:10.1371/journal.pone.0107042.

Tatem, A.J., (2014). Mapping the denominator: Spatial demography in the measurement of progress. *International Health*, vol. 6, No. 3, pp. 153-155. Available from doi:10.1093/inthealth/ihu057.

Tatem, A. J., and others (2014). Integrating rapid risk mapping and mobile phone call record data for strategic malaria elimination planning. *Malaria Journal*, vol. 13, No. 52.

_____ (2011). The effects of spatial population dataset choice on estimates of population at risk of disease. *Population Health Metrics*, vol. 9, No. 4. Available from doi:10.1186/1478-7954-9-4.

_____ (2013). Millennium development health metrics: where do Africa's children and women of childbearing age live? *Population Health Metrics*, vol. 11, No. 1. Available from doi:10.1186/1478-7954-11-11.