Using socio-economic well-being to predict geospatial epidemic intensity in a developing country setting

Rahman Sanya
Makerere University
P. O. Box 7062
Kampala, Uganda
hbasanya@gmail.com

Ernest Mwebaze
Makerere University
P. O. Box 7062
Kampala, Uganda
emwebaze@cit.ac.ug

Abstract

Unavailability and poor quality data on disease occurence often is major hinderance to proper healthcare delivery in developing countries. We use data on socio-economic well-being to predict spatial disease distribution patterns in a developing country setting. Specifically, we use spatial poverty data as proxy to estimate disease cases at lower spatial units basing on data at a higher spatial unit. Being a first version of the model prediction accuracy and correlation between observed and predicted values are low. With further optimization and incorporation of other known disease risk factors we think the performance of our model will improve.

Keywords: socio-economic well-being, infectious diseases, spatial predictive modeling, developing country.

1 Introduction

For a long time infectious disease outbreaks have inflicted huge socio-economic and health burdens on developing countries due to their weak healthcare systems. E.g., the United Nations Development Program (2015) estimated the recent Ebola outbreak in West Africa to cost each of directly affected countries millions of US dollars. In Uganda, one of African countries worst hit by HIV/AIDS, it was forecast the country's economy would shrink by 39% by 2025 due to effects of the disease (Ministry of Finance Uganda, 2008).

A major hindrance to fighting disease outbreaks in developing countries is absence/poor quality data on disease prevalence in populations. To address this problem some sub-Saharan African countries setup Health Management Information Systems (HMIS). A HMIS is a countrywide integrated reporting system for collecting, transmitting, and storing high quality routine data on healthcare services and disease occurrence from a health facility level to national level. Data from HMIS is meant for, among other uses, detecting and predicting epidemics (Ministry of Health Uganda, 2010). Despite setting up HMIS, the problem of unavailability of data has persisted. E.g. Ohiri et al. (2016) found problem of limited malaria data in District Health Information Systems (DHIS) in Nigeria. Empirical analyses of malaria records by Andrade-Pacheco et al. (2014) and TB and HIV/AIDS records (this study) obtained from Uganda's HMIS revealed gaps due to non-reporting of case counts over spatio-temporal dimensions. Lack of consistent reporting of disease case counts gives rise to inaccurate records at various levels of reporting with consequence that intervention planning is negatively impacted and hence, failure to effectively counter epidemics.

There are different ways to deal with the problem of data unavailability on disease occurrence in a spatial region. A summary of such methods is found in Clements et al. (2013). Among them are geostatical and Bayesian methods for spatial data analysis and interpolation. E.g., a model-based geostatistical framework were used to produce high-resolution poverty maps by Tatem et al. (2014) and malaria risk maps by Hay et al. (2009). Generalized linear models and Gaussian process regression were used to predict malaria incidence by Mubangizi et al. (2014) and Bhatt et al. (2017). While these techniques are currently state-of-the-art, they require significant knowledge of statistics making them inaccessible to non-experts.

In this paper a simple model is proposed for estimating disease case counts over a spatial region. This is done by spatially disaggregating case numbers at higher administrative units to lower administrative units based on socio-economic characteristics of a region. Specifically, we use poverty data to downscale case counts at higher administrative units to corresponding lower administrative units by first downscaling the data to grids of $100m^2$ and then summing per grid counts to provide numbers at the next higher administrative units. The method exploits association that is reported to exist between socio-economic well-being and infectious diseases in developing country settings.

The rest of this paper is organized as follows, section 2 points out some recent literature, section 3 outlines methods used while results are in section 4. We suggest future work in section 5 and conclude in section 6.

2 Related work

Previous studies have examined the relationship between living conditions and disease. E.g., Kirenga et al. (2015) investigated TB risk factors in adults and identified poverty as a leading cause, while Tusting et al. (2016) found association between poverty and malaria prevalence in children. Both studies conducted in Uganda were at level of individual persons, involved small study areas, and had no spatial dimension to their data/analysis. In the case of non-infectious diseases, Ludwing et al. (2012) reported link between neighborhood characteristics and physical/mental health. Using a multi-level analysis framework association is reported between neighborhood greenness and chronic health conditions (Brown et al., 2016). The latter two studies were conducted in the U.S.A.

3 Data and methods

3.1 Datasets

3.1.1 Poverty data

The poverty dataset was downloaded from WorldPop website (http://www.worldpop.org.uk). It consists of predicted proportion of ‘poor' people per 1km2 grid for year 2015. We used dataset on multi-dimensional poverty index (MPI) which considers broad definition of well-being (Alkire & Santos, 2010). The data in raster file format was resampled from $1km^2$ to $100m^2$ to be consistent with other datasets used in a broader study. Statistics were computed on each grid to serve as poverty index that were then used to weight each grid (Eq. 2).

3.1.2 Disease case notification

The disease we use as case study is Tuberculosis (TB), a high burden infectious disease and a leading cause of death among HIV/AIDS patients in Uganda (World Health Organization, 2017; UNAIDS, 2013). The dataset consisting of monthly case counts for year 2015 were acquired from HMIS, Ministry of Health Uganda.

3.2 Methods

3.2.1 Study area

The area of study is central/northern regions of Uganda. Uganda lies between 10 29' South and 40 12' North latitude, 290 34' East and 350 0' East longitude. It has population of 34.6 million people and covers area of $241{,}551km^2$ (Uganda Bureau of Statistics, 2016). The spatial units of analysis we used are local administrative unit 1 (LAU1) (district) and LAU3 (subcounty). These were selected to correspond with disease datasets. A map of the study area is shown in Fig. 1.

3.2.1 Estimating disease case numbers

We used two methods to estimate case counts at LAU3 based on counts at LAU1. In first method we disaggregated case numbers at LAU1 to $100m^2$ grids weighted by spatial poverty index. The case counts for all grids situated within boundary of a LAU3 unit are then summed to provide predicted count for that LAU3. Since we did not have case counts for LAU1 units we first calculated those by aggregating observed counts at LAU3 units at time *t* using Eq. (1),

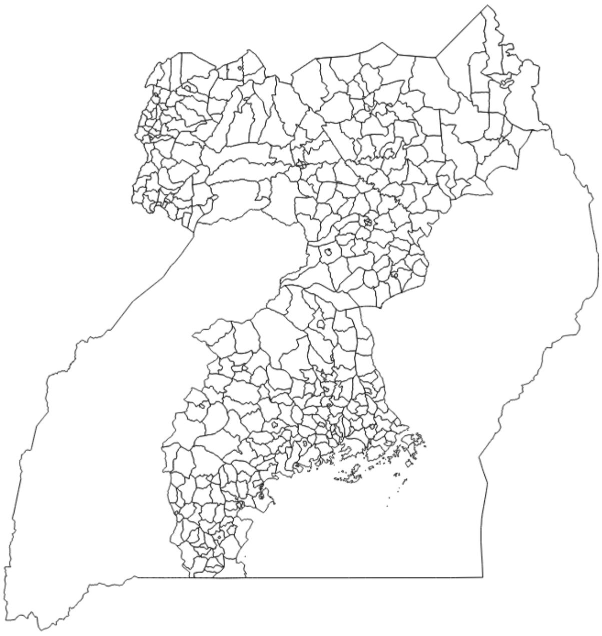
$$C_d = \sum_{j=1}^{n} y_j \qquad (1)$$

where y_j is observed case count at a LAU3 unit and C_d is case number at a LAU1 unit.

Each pixel's weight w_i was calculated based on spatial poverty index *v* using Eq. (2),

$$w_i = \frac{v_i}{\sum v_i} \qquad (2)$$

Figure 1: Map of Uganda showing study area (central-northern regions) delineated by LAU3.



Disease case count per grid c_i was then calculated based on grid weight *w*as, Eq. (3),

$$c_i = C_d * w_i \qquad (3)$$

The predicted count per pixel $\hat{y}_j$ for a LAU3 unit *j* was got by summing estimated case counts for all grids falling within boundary of that LAU3 unit as shown in Eq. (4),

$$\hat{y}_j = \sum_j c_i \qquad (4)$$

In the second method, a baseline model is used where case count at a LAU1 unit is downscaled directly to LAU3 units weighted by grid count (Eq. 5),

$$\hat{y}_j = \frac{C_d * n_j}{N_d} \qquad (5)$$

where n_j is the count of $100m^2$ grids in LAU3 unit *j* and N_d is total grid count in LAU1 *d*, respectively.

The observed and predicted case numbers were normalized by dividing by population of each LAU3 unit to obtain per capita count before performing analysis.

We evaluated the models using standard validation statistics including mean absolute error (MAE) to test prediction accuracy and Pearson correlation coefficient *r* for correlation between observed and predicted case numbers. The MAE was selected for its simplicity and ease of interpretation. Further evaluation was conducted to examine spatial non-stationarity using geographically weighted regression (GWR) (Fotheringham, Brunsdon & Charlton, 2002). Use of GWR also helps to address problem of 'scale effect'.

4 Results and discussion

4.1 Results

4.1.1 Understanding the disease dataset

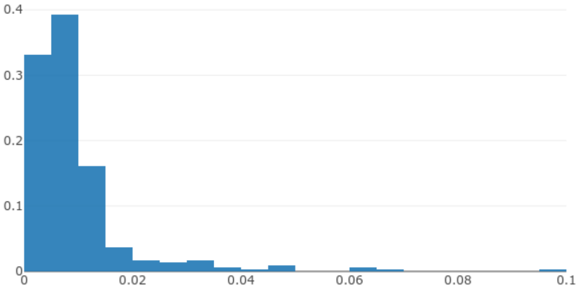
Exploratory analysis was carried out to understand the basic statistical properties of our disease dataset. Summary statistics are provided in Table 1.

Table 1: Statistical characteristics of per capita TB datasets.

	TB (observed)	TB (poverty)	TB (grid)
Mean	0.009670	0.009838	0.010011
Median	0.006684	0.006247	0.006403
Mode	0.005377	0.010091	0.009810

Both observed and predicted datasets exhibit non-normal right-skewed distribution, Fig. 2 and 3. The datasets also have outliers above maximum values consistent with right-skewed distribution, Fig. 4. A possible explanation for this pattern is that while majority of LAU3 units (rural units) have low per capita TB prevalence, a few units (urban units) have high prevalence.

Figure 2: Histogram plot for observed per capita TB dataset exhibiting non-normal, skewed-right distribution.



4.1.2 Accuracy and correlation assessment

MAE and *r* values are shown in Table 2. The two models have comparable MAE and *r* values. Though generally low, the correlation seems promising given that this is a first version of our model without any optimization being applied.

4.1.3 Visualizing spatial epidemic intensity

Choropleth maps were created using a Geographic Information System (GIS) tool (QGIS version 2.18.13). As seen in Fig. 5, 6, and 7 our models tend to over estimate per capita TB across all but a handful of LAU3 units. This is moreso for the model based on spatial poverty distribution.

Figure 3: Histogram plot for poverty-predicted per capita TB dataset exhibiting non-normal, skewed-right distribution.

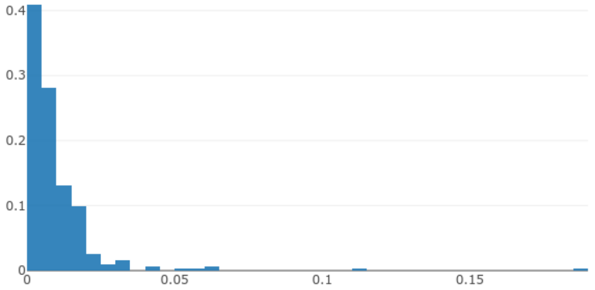


Figure 4: Box plots for per capita TB datasets. All three datasets exhibit similar out-lier distribution pattern.

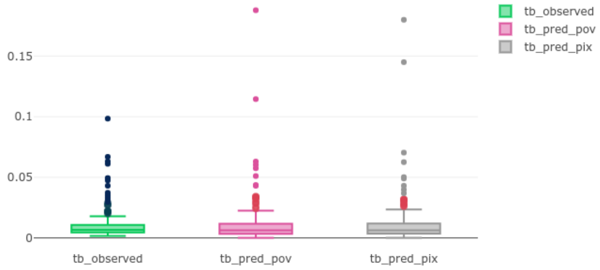


Table 2: Results of accuracy and correlation analyses on observed and predicted per capita TB.

Measure	TB (poverty)	TB (grid count)
MAE	0.007893	0.007715
Pearson *r*	0.275186	0.338664
p-value	1.045727e-06	1.197466e-09

4.1.4 Spatial relationship analysis

We used R (version 3.4.4) to run a GW regression to further assess the relationship between poverty and TB based on a Gaussian model using an adaptive kernel with bandwidth (0.4150209) calibrated by cross-validation method. A map of local coefficients is shown in Fig. 8. The local coefficients vary from -0.015575 to 0.009121, a range within which the global regression coefficient (-0.010429) falls.

4.2 Discussion

The goal of this work is to quantify effect of a socio-economic factor on spatial dispersal of infectious diseases. We have partially investigated how poverty data may be used to predict spatial TB distribution. The model however, over-estimates TB in rural areas while under-estimating it in urban areas. This is in contrast to current situation where TB is more prevalent in urban areas than in rural areas (Ministry of Health Uganda, 2017).

One reason our model performs poorly could be due to fact it does not account for spatial heterogeneity in TB prevalence mentioned above. In addition, several risk factors are known to be associated with infectious diseases in developing countries. E.g., Kirenga et al. (2015) found that TB in Kampala is associated with several risk factors including HIV/AIDS, overcrowding, and alcohol use. We have not incorporated potential effect of these factors into our model. We also acknowledge other limitations of this work including use of small dataset and possible bias in poverty and/or disease datasets that we may not be aware of. The model has also not been evaluated on datasets spanning more than one year.

Figure 5: Map of per capita TB (observed) by local administrative unit 3 (LAU3) in study area for 2015.

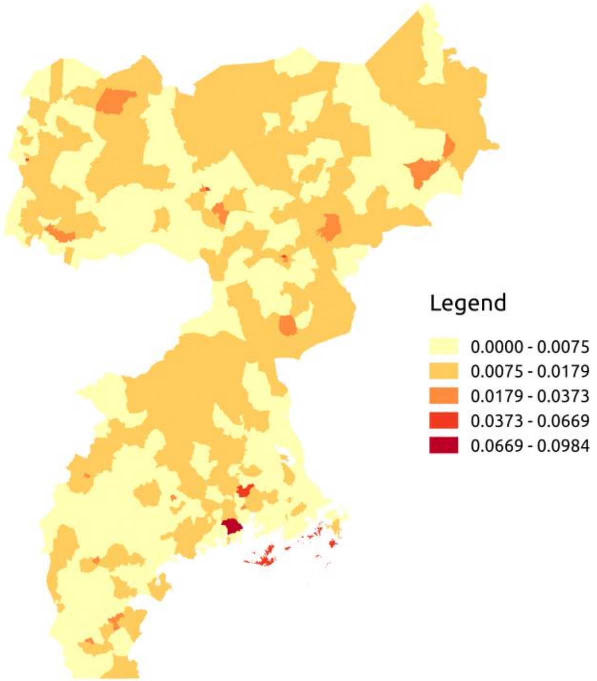
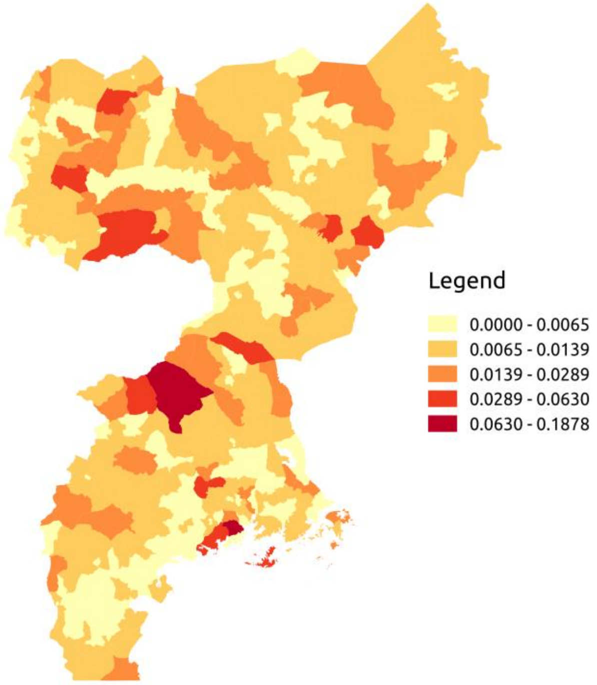


Figure 6: Map of per capita TB (poverty-based prediction) by local administrative unit 3 (LAU3) in study area for 2015.



Although current prediction accuracy is low, we plan to build on these results to improve performance of our model by optimizing it using parameters learnt from training data. A potential method to address heterogeneity in spatial infection distribution is to apply a superlinear parameter to predicted case numbers in urban areas as suggested for such phenomena by Schlapfer et al. (2014) using Eq. (6).

$$y_j = \alpha\, \hat{y}_j^{\beta} \qquad (6)$$

where $y_j=\left[y_{j1,}y_{j2,\ldots,}y_{jn}\right]$ and

$\hat{y}_j=\left[\hat{y}_{j1,}\hat{y}_{j2,\ldots,}\hat{y}_{jn}\right]$.

The parameter α is a scaling ratio while β accounts for superliner effect of population density on epidemic intensity in an urban area.

Apart from the fact that our study and Brown et al. (2016) address different categories of diseases under different settings, there are other more important differences. E.g., while both studies use datasets aggregated at different spatial scales the latter study integrates all data into one multi-level statistical framework before analyzing it. In our case however, processing and analysis is performed at each level using a non-statistical approach. Secondly, while their study sought to establish a relationship between health outcomes and physical environment (i.e. vegetation), the aim of our study is to use such a relationship to predict spatial disease distribution.

Figure 7: Map of per capita TB (grid count-based prediction) by local administrative unit 3 (LAU3) in study area for 2015.

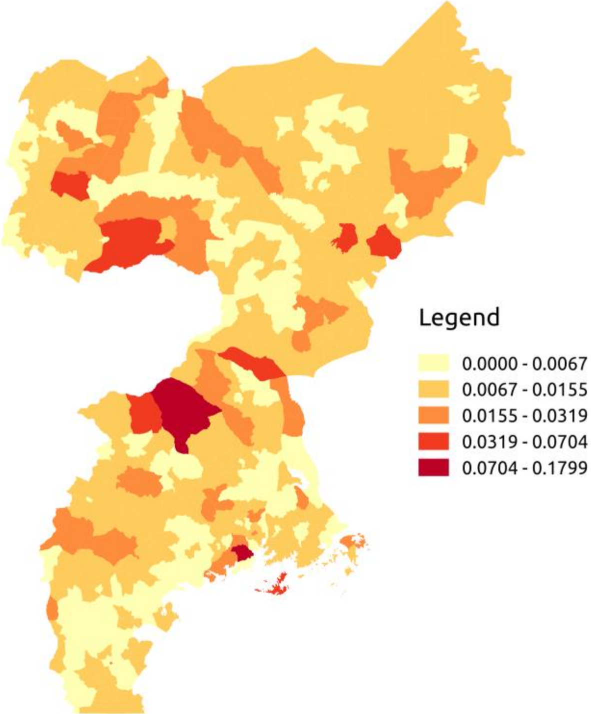
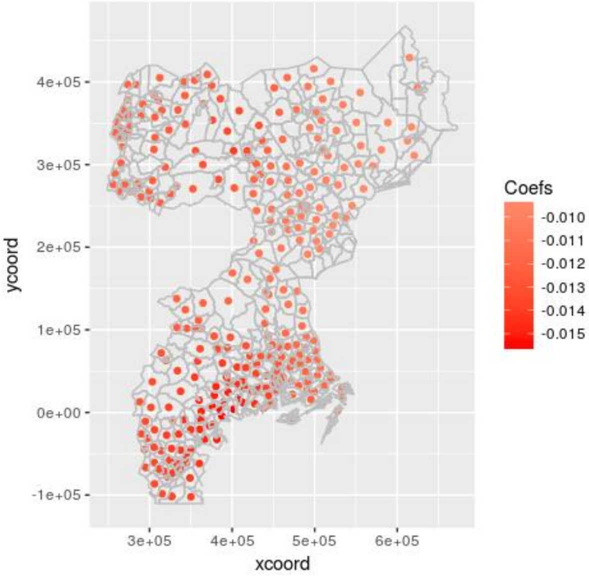


Figure 8: Map of GWR coefficients over study area



5 Future work

A limitation of our model is treating spatial units as 'closed systems' that do not interact with each other with potential implications for epidemic. It has been demonstrated elsewhere that population dynamics (e.g mobility) have influence on spatial disease distribution. For this reason, we plan to integrate population mobility into our model.

In context of developing countries, heterogeneity in socio-economic well-being are known to occur within even small spatial units than what we have considered here. It might therefore, be unreasonable to assume uniform case distribution over such relatively large spatial units. To address this, we plan to use smaller spatial units for analysis.

We also intend to incorporate effect of other risk factors into the model including access to health care, HIV/AIDS prevalence, and housing conditions. Accounting for higher prevalence of disease in urban areas is also another research direction we intend to pursue.

6 Conclusion

We report early results of an attempt to use poverty data to estimate infectious disease numbers over smaller spatial units by disaggregating numbers at larger units. This being a first, un-optimized version of our model the prediction accuracy is still low which we plan to build on.

Acknowledgments R. S would like to thank the AFDB-HEST Project for financial assistance.

References

Alkire, S. & Santos, M. (2010) *Acute multidimensional poverty: A new index for developing countries*. UNDP-HDRO, New York.

Andrade-Pacheco, R., Mubangizi, M., Quinn, J. & Lawrence, N. D. (2014) Consistent mapping of government malaria records across a changing territory delimitation. *Malaria Journal*, 13(Suppl.):5.

Bhatt, S., Cameroon, E., Flaxman, S. R., Weiss, D. J., Smith, D. L., & Gething, P. W. (2017) Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *Journal of Royal Society Interface*, 14:20170520.

Brown, S. C., Lombard, J., Wang, K., Byrne, M. M., Toro, M., Plater-Zyberk, E., Feaster, D. J., Kardys, J., Nardi, M. I., Perez-Gomez, G., Pantin, M. H., & Szapocznik, J. (2016) Neighborhood greenness and chronic health conditions in Medicare beneficiaries. *American Journal of Preventive Medicine*, 51(1), 78-89.

Clements, A. C. A., Reid, H. L., Kelly, G. C., & Hay, S. I. (2013) Further shrinking the malaria map: how can geospatial science help to achieve malaria elimination?. *Lancet Infect Dis.*, 13, 709-718.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. *Geographically Weighted Regression: The analysis of spatially varying relationships*. Wiley, 2002.

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. *PloS Medicine*, 2286-0302, DOI 10.1371/journal.pmed.1000048.

Kirenga, B. J., Ssengooba, W., Muwonge, C., Nakiyingi, L., Kyaligonza, S., Kasozi, S., Mugabe, F., Boeree, M., Joloba, M. & Okwera, A. (2015) Tuberculosis risk factors among tuberculosis patients in Kampala, Uganda: Implications for tuberculosis control. *BMC Public Health*, 15(13), 1-7, DOI 10.1186/s12889-015-1376-3.

Ludwig, J., Duncan, G. J., Gennetian, L. A., Lwarence, F. K., Kessler, R. C., Kling, J. R., & Sanbonmatsu, L. (2012) Neighborhood effects on the long-term well-being of low-income adults. *Science*, 337(6101), 1505-1510.

Ministry of Finance Uganda. (2008) *Assessing the Macro-economic Impact of HIV/AIDS in Uganda*. Summary Report.

Ministry of Health Uganda. (2010) *The Health Management Information System*. Volume 1. Health Unit Procedure Manual.

Uganda Ministry of Health. (2017) *The Uganda National Tuberculosis Prevalence Survey*, 2014-2015.

Mubangizi, M., Andrade-Pacheco, R., Smith, M., Quinn, J. A., & Lawrence, N. (2014) Malaria surveillance with multiple data sources using Gaussian process models. In: *Proceedings of the 1st International Conference on the Use of Mobile ICT in Africa 2014*, Stellenbosch South Africa, 2014.

Ohiri, K., Ukoha, N. K., Nwangu, C. W., Chima, C. C., Ogundeji, Y.K., Rone, A., & Reich, M. R. (2016) An assessment of data availability, quality, and use in malaria program decision making in Nigeria. *Health Systems & Reform*, 2(4), 319-330. DOI 10.1080/23288604.2016.1234864.

Schlapfer, M., Bettencourt, L. M. A., Grauwin, S., Raschke, M., Claxton, R., Smoreda, Z., West, G. B., & Ratti, C. (2014) The scaling of human interactions with city size. *Journal of Royal Society Interface*, 11(98:20130789), DOI 11(98:20130789).

Tatem, A. J., Gething, P. W., Bhatt, S. W. D., & Pezzulo, C. (2014) *Final report: Development of high-resolution gridded poverty surfaces*. University of Southampton and University of Oxford.

Tusting, L. S., Rek, J., Arinaitwe, E., Staedke, S. G., Kamya, M. R., Cano, J., Bottomley, C., Johnston, D., Dorsey, G., Lindsay, S. W., & Lines, J. (2016) Why is malaria associated with poverty? Findings from a cohort study in rural Uganda. *Infectious Diseases of Poverty*, 5(78). DOI 10.1186/s40249-016-0164-3.

Uganda Bureau of Statistics. (2016) *The National Population and Housing Census 2014 – Main Report*. Kampala, Uganda.

The UNAIDS. (2013) *Global Report: UN AIDS Report on the Global AIDS Epidemic 2013*.

The United Nations Development Program. (2015) *Socio-economic Impact of Ebola virus disease in West African Countries: A Call for National and Regional Containment, Recovery and Prevention*.

The World Health Organization. (2017) *Global Tuberculosis Report 2017*.