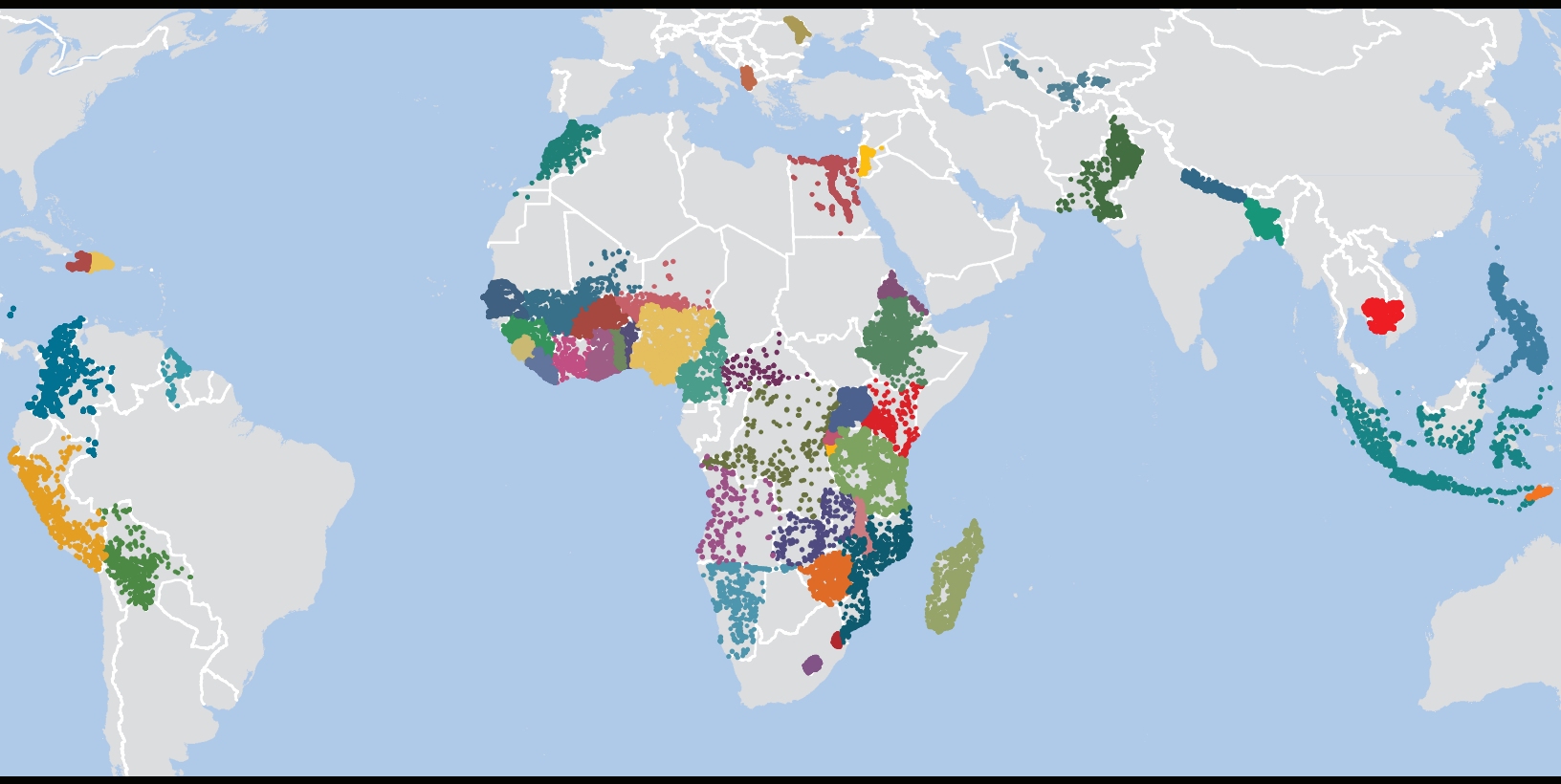
USAID
FROM THE AMERICAN PEOPLE

CREATING SPATIAL INTERPOLATION SURFACES WITH DHS DATA

DHS SPATIAL ANALYSIS REPORTS 11



SEPTEMBER 2015
This publication was produced for review by the United States Agency for International Development (USAID). The report was prepared by Peter Gething, Andy Tatem, Tom Bird, and Clara R. Burgert-Brucker of ICF International, Rockville, MD, USA.

DHS Spatial Analysis Reports No. 11

Creating Spatial Interpolation Surfaces with DHS Data

Peter Gething

Andy Tatem

Tom Bird

Clara R. Burgert-Brucker

ICF International

Rockville, Maryland, USA

September 2015

Corresponding author: Clara R. Burgert-Brucker, International Health and Development, ICF International, 530 Gaither Road, Suite 500, Rockville, MD 20850, USA; phone: +1-301-572-0446; fax: +1-301-407-6501; email: clara.burgert@icfi.com

Acknowledgment: The authors would like to acknowledge Shireen Assaf, Mike Emch, and Mahmoud Elkasabi for their review of this report.

Editor : Diane Stoy
Document Production: Natalie La Roche

This study was carried out with support provided by the United States Agency for International Development (USAID) through The DHS Program (#AID-OAA-C-13-00095). The views expressed are those of the authors and do not necessarily reflect the views of USAID or the United States Government.

The DHS Program assists countries worldwide in the collection and use of data to monitor and evaluate population, health, and nutrition programs. For additional information about The DHS Program contact: DHS Program, ICF International, 530 Gaither Road, Suite 500, Rockville, MD 20850, USA. Phone: +1-301-407-6500; fax: +1-301-407-6501; email: reports@dhsprogram.com; Internet: www.dhsprogram.com.

Recommended citation:

Gething, Peter, Andy Tatem, Tom Bird, and Clara R. Burgert-Brucker. 2015. *Creating Spatial Interpolation Surfaces with DHS Data* DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International.

Contents

Tables v

Figures v

Abbreviations ix

Preface xi

Abstract xiii

Executive Summary xv

1. Background and Objectives 1
   - 1.1 Objective 1: Exploring the Potential of Bayesian Geostatistics for Interpolating DHS Survey Data 1
   - 1.2 Objective 2: Exploring the Impact of DHS Cluster Displacement on the Production of Interpolated Surfaces 2
   - 1.3 Objective 3: Investigating the Potential for Novel Methodologies and Covariates to Address the Challenge of Mapping within Urban Areas 3

2. Selection of DHS Indicators and Exemplar Countries 5

3. National-Level Geostatistical Mapping 7
   - 3.1 Methods 7
     - 3.1.1 Overview 7
     - 3.1.2 Preparation of Georeferenced DHS Indicator Data 7
     - 3.1.3 Exploratory Analysis 11
     - 3.1.4 Geostatistical Modeling 13
   - 3.2 Results 14
     - 3.2.1 Access to HIV Testing in Women 15
     - 3.2.2 Stunting in Children 21
     - 3.2.3 Anemia Prevalence in Children 26
     - 3.2.4 Access to Improved Sanitation 31

4. Effects of Cluster Centroid Displacement 37
   - 4.1 Methods 37
     - 4.1.1 Overview 37
     - 4.1.2 Generating Displaced Data 37
     - 4.1.3 Exploration of Effect of Displacement on Statistical Properties of Data 39
     - 4.1.4 Exploration of Effect of Displacement on MBG-derived Interpolated Surfaces 39
   - 4.2 Results 40
     - 4.2.1 Generating Displaced Data 40
     - 4.2.2 Exploration of Effect of Displacement on Statistical Properties of Data 44
     - 4.2.3 Exploration of Effect of Displacement on MBG-derived Interpolated Surfaces 46

5. The Potential of High Resolution Urban Mapping 59
5.1 Methods 59
5.1.1 Urban Subsets of the Data and Urban Definitions 59
5.1.2 High Resolution Urban Covariates 59
5.1.3 Effects of Displacement on Linear Models in Urban Areas 60
5.1.4 Exploration of Different Approaches to Modeling Urban Settings at High Resolutions 61
5.1.5 Kano Urban Dataset 61
5.2 Results 64
5.2.1 Urban Definitions 64
5.2.2 Effects of Displacement on Model Estimates 66
5.3 Exploration of Model Approaches to High-Resolution Datasets 73
5.3.1 Comparison of Generalized Additive Models (GAM), Boosted Regression Trees (BRT) and Linear Models (LM) for Predicting Proportion of Children under 5 Years Old 73
5.3.2 Effects of Displacement on Models at Differing Resolutions 75

6. Discussion 77
6.1 National-level Geostatistical Mapping 77
6.2 Effects of Cluster Centroid Displacement 78
6.3 The Potential of High Resolution Urban Mapping 78
6.4 Next Steps 79
6.4.1 Further Improvements to the Methodology 79
6.4.2 Development of Use-cases for Predictive Maps 81

7. Conclusion 83

References 85

Tables

Table 1. Details of geospatial covariates assembled for use in mapping 9
Table 2. Summary output from the covariate selection procedure for the indicator on proportion of women accessing HIV testing in the 12 months preceding the survey 17
Table 3. Summary validation results from the model-based geostatistical map for the indicator on proportion of women accessing HIV testing in the 12 months preceding the survey 17
Table 4. Summary output from the covariate selection procedure for the indicator on proportion of children <5 years who are stunted 22
Table 5. Summary validation results from the model-based geostatistical map for the indicator on proportion of children 6-59 months who are stunted 22
Table 6. Summary output from the covariate selection procedure for the indicator on proportion of children 6-59 months who are anemic 27
Table 7. Summary validation results from the model-based geostatistical map for the indicator on proportion of children 6-59 months who are anemic 27
Table 8. Summary output from the covariate selection procedure for the indicator on proportion of households with improved sanitation. PR2 is the predictive R-squared statistic 32
Table 9. Summary validation results from the model-based geostatistical map for the indicator on proportion of households with improved sanitation 32
Table 10. Summary of relative influence for different covariates in the boosted regression tree model for proportion of children under 5 74
Table 11. Summary of the parameters for different covariates in the full linear model (LM) 74
Table 12. Summary of the relative influence of different covariates in the Generalized Additive Model at the household level 75

Figures

Figure 1. Example of a typical variogram showing the different features as described in the text 12
Figure 2. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of women accessing HIV testing in the 12 months preceding the survey .. 16
Figure 3. (left) Predicted map of indicator on proportion of women accessing HIV testing the 12 months preceding the survey in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 18
Figure 4. (left) Predicted map of indicator on proportion of women accessing HIV testing the 12 months preceding the survey in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 19
Figure 5. (left) Predicted map of indicator on proportion of women accessing HIV testing the 12 months preceding the survey in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 20
Figure 6. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of children 6-59 months who are stunted 21
Figure 7. (left) Predicted map of indicator on proportion of children 6-59 months who are stunted in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 23
Figure 8. (left) Predicted map of indicator on proportion of children 6-59 months who are stunted in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 24

Figure 9. (left) Predicted map of indicator on proportion of children 6-59 months who are stunted in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 25
Figure 10. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of children 6-59 months who are anemic 26
Figure 11. (left) Predicted map of indicator on proportion of children 6-59 months who are anemic in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 28
Figure 12. (left) Predicted map of indicator on proportion of children 6-59 months who are anemic in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 29
Figure 13. (left) Predicted map of indicator on on proportion of children 6-59 months who are anemic in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 30
Figure 14. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of households with improved sanitation 31
Figure 15. (left) Predicted map of the indicator on proportion of households with improved sanitation in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 33
Figure 16. (left) Predicted map of the indicator on proportion of households with improved sanitation in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 34
Figure 17. (left) Predicted map of the indicator on proportion of households with improved sanitation in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty 35
Figure 18. Example set of randomly displaced points implemented according to the standardized DHS displacement algorithm using the 2010 Tanzania DHS 38
Figure 19. Location of the 100 randomly displaced cluster centroid from the 2008 Ghana DHS 41
Figure 20. Location of the 100 randomly displaced cluster centroid from the 2010 Tanzania DHS 42
Figure 21. Location of the 100 randomly displaced cluster centroid from the 2011 Uganda DHS 43
Figure 22. The effects of centroid displacement on empirical variograms structure 45
Figure 23. The effects of centroid displacement on indicator relationships with gridded covariates 47
Figure 24. Effect of displacement on performance of model-based geostatistical predictive performance for the HIV testing indicator 49
Figure 25. Effect of displacement on performance of model-based geostatistical predictive performance for the stunting indicator 50
Figure 26. Effect of displacement on performance of model-based geostatistical predictive performance for the anemia indicator 51
Figure 27. Effect of displacement on performance of model-based geostatistical predictive performance for the sanitation indicator 52
Figure 28. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS 54
Figure 29. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS 55
Figure 30. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS 56
Figure 31. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS 57

Figure 32. Plots of household survey points (A) and gridded (25 m) PU5 data (B) for the microcensus surveys in the Gama ward of Kano, Nigeria. 62
Figure 33. Clusters represented by bars, with heights proportional to log(population density) and ordered by population density 65
Figure 34. Urban clusters as defined by the DHS classification represented by bars, with heights proportional to log(population density) and ordered by population density rank 66
Figure 35. Effects of displacement and buffering on diagnostics for models on the prevalence of stunting in children in Tanzania 67
Figure 36. Effects of displacement and buffering on diagnostics for models on the prevalence of stunting in children in Uganda 68
Figure 37. Effects of displacement and buffering on diagnostics for models on the prevalence of anemia in children in Tanzania 69
Figure 38. Effects of displacement and buffering on diagnostics for models on the prevalence of anemia in children in Tanzania 70
Figure 39. Effects of displacement and buffering on diagnostics for models on the prevalence of anemia in children in Ghana 71
Figure 40. Effects of displacement and buffering on diagnostics for models on the prevalence of improved sanitation in Tanzania 72
Figure 41. Effects of displacement and buffering on diagnostics for models on the prevalence of improved sanitation in Uganda 73
Figure 42. Summary of model diagnostics (PR2) for a series of 50 models run on displaced data in urban areas 76

Abbreviations

AIC	Akaike information criterion
AIS	AIDS Indicator Survey
BRT	Boosdted Regression Trees
DHS	Demographic and Health Survey
EA	Enumeration Area
EU	European Union
GAM	Generalized Additive Models
GIS	Geographic Information System
GPS	Global Positioning System
HIV	Human Immunodeficiency Virus
INLA	Integrated Nested Laplace Approximations
LM	Linear Models
MAE	Mean Absolute Error
MBG	Model-Based Geostatistics
MIS	Malaria Indicator Survey
PSU	Primary Sampling Unit
RMSE	Root-Mean-Square Error
SD	Standard Deviation
USAID	United States Agency for International Development
WHO	World Health Organization

Preface

The Demographic and Health Surveys (DHS) Program is one of the principal sources of international data on fertility, family planning, maternal and child health, nutrition, mortality, environmental health, HIV/AIDS, malaria, and provision of health services.

The DHS Spatial Analysis Reports supplement the other series of DHS reports to meet the increasing interest in a spatial perspective on demographic and health data. The principal objectives of all DHS report series are to provide information for policy formulation at the international level and to examine individual country results in an international context.

The topics in the DHS Spatial Analysis Reports are selected by The DHS Program in consultation with the U.S. Agency for International Development. A range of methodologies are used, including geostatistical and multivariate statistical techniques.

It is hoped that the DHS Spatial Analysis Reports series will be useful to researchers, policymakers, and survey specialists, particularly those engaged in work in low- and middle-income countries, and will be used to enhance the quality and analysis of survey data.

Sunita Kishor
Director, The DHS Program

Abstract

Improved understanding of sub-national geographic variation and inequity in demographic and health indicators is increasingly recognized as central to meeting development goals. Data from DHS surveys are critical to monitoring progress in these indicators but are generally not used to support sub-national evaluation below the first-level administrative unit. This study explored the potential of geostatistical approaches for the production of interpolated surfaces from GPS cluster located survey data, and for the prediction of gridded surfaces at 5×5km resolution. The impact of DHS cluster displacement on these interpolated surfaces and the particular challenges of mapping in highly heterogeneous urban areas were also investigated.

A flexible and robust geostatistical framework was proposed for generating and validating interpolated surfaces with georeferenced DHS survey data. The framework was tested by using four indicators from three different surveys. The accuracy of the interpolated surfaces varied between settings, and was driven by the spatial structure of each indicator and its relationship to available covariate data.

The random displacement of DHS cluster geopositioning information reduced the precision of predicted maps, although the impact varied between settings and was generally modest. Over shorter distances, the greater degree of geographical heterogeneity that was associated with urban areas meant they were more sensitive to the impact of cluster displacement. High-resolution covariates and novel statistical approaches showed potential for improving mapping in these areas.

This study demonstrated that, with appropriate modeling and validation, such data have broad utility for creating maps of a wide range of indicators that support improved geographically stratified decisionmaking.

Executive Summary

Background and Objectives

Improved understanding of sub-national geographic variation and inequity in health status, wealth, and access to resources within countries is increasingly recognized as central to meeting development goals. Monitoring demographic, access, and health status inequalities for targeting interventions, and measuring progress towards health and development goals require a reliable and detailed evidence base. Comprehensive, contemporary, and spatially detailed information on demographic and health attributes of populations currently remain usable only at aggregated regional levels through national household surveys such as those conducted by The Demographic Health Survey (DHS) Program. The DHS Program has been a leader in collecting and providing cluster-randomized survey data on core development indicators. The availability of the Global Positioning System (GPS) coordinates for DHS (and other survey program) clusters provides local scale information for quantifying demographic and health status heterogeneities and inequities.

Many recent studies have utilized a range of interpolation approaches to derive continuous estimated surfaces of variables of interest from DHS household surveys. This indicates a demand for interpolated maps created from survey locations to calculate the estimates for areas that are smaller than those currently provided by The DHS Program. How these maps should be produced; whether reliable, accurate maps are feasible in both rural and urban geographies; how data displacement influences anonymity; what form the outputs should take; and which variables should be mapped remain questions to be answered. This report addresses three overarching aims:

1. To explore the potential of Bayesian model-based geostatistics for the production of interpolated surfaces from GPS cluster located survey data.
2. To explore the impact of DHS cluster displacement on the production of interpolated surfaces.
3. To investigate the potential for novel methodologies and covariates to address the challenge of mapping within urban areas.

The Potential of Bayesian Model-based Geostatistics for the Production of Interpolated Surfaces from GPS Cluster Located Survey Data

Three recent DHS surveys in sub-Saharan Africa were used for the analyses: the 2008 Ghana DHS, 2010 Tanzania DHS, and the 2011 Uganda DHS. Four indicators were identified from each, and were used for all subsequent analyses which included access to HIV testing in women, prevalence of stunting in children, prevalence of anemia in children, and household access to improved sanitation.

For each of the 12 survey indicators, a Bayesian geostatistical model was built and validated to generate an interpolated surface. Bayesian geostatistical models take traditional geostatistical approaches to interpolation such as *kriging*, and provide a more formal approach to identifying and propagating uncertainty because the model is fitted to the data. For every pixel on the mapped surface, the full model output is a posterior distribution for the predicted variable; this represents a complete model of the uncertainty around the estimated value. These can be summarized by using a point estimate (such as the posterior mean) to generate a mapped surface as well as a corresponding map of uncertainty.

We were able to generate plausible surfaces for each of the 12 survey indicator combinations. The validation statistics and uncertainty maps demonstrated that model performance was acceptable in all cases, although the level of precision varied between settings. This was driven by the amount and spatial scale of variation

displayed by each indicator, and the extent to which the indictor was linked to the available geospatial covariates. This means that it is unlikely to be possible to provide prior assessment of the feasibility of mapping particular indicators in particular countries. However, the methodological framework will yield maps with appropriate uncertainty metrics, on which decisions can subsequently be made on the required precision for the intended use of the map.

The Impact of DHS Cluster Displacement on the Production of Interpolated Surfaces

All DHS survey data are subject to a standardized displacement procedure that protects respondents' anonymity; this is especially important in the more sensitive aspects of the questionnaires such as those that inquire about HIV infection status. To test the impact of this displacement on geostatistical mapping, we took the non-displaced dataset for each survey indicator and generated 100 sets of displaced data by following the standard DHS protocol. With this experimental dataset, we implemented 100 geostatistical models for each survey indicator, conducted a validation exercise for each, and compared model performance with the reference non-displaced set.

As expected, the results were complex with the extent of the impact of displacement varying between indicators and between the individual survey data sets. Nevertheless, a number of overarching conclusions can be drawn. First, the impact of displacement on summary validation statistics (i.e., the overall "performance" of a geostatistical model) modest, although the effect was more marked where the non-displaced model was relatively high performing. Second, the impact of displacement varied geographically across each interpolated surface. This was apparently driven by several factors such as data point density and the heterogeneity of important geographic covariates.

The Potential for Novel Methodologies and Covariates to Address the Challenge of Mapping within Urban Areas

Urban areas present a particular challenge to interpolation because variation in indicators is often pronounced over short distances. This reflects the typically patchy, heterogeneous nature of urban landscapes. In this study, the potential of much higher resolution satellite and GIS-derived covariates that capture urban details were tested for their abilities to improve mapping accuracies. Moreover, novel approaches were developed by using a case study of mapping proportions of the population under 5 years of age in Kano City, Nigeria. A combination of high resolution covariates and non-linear techniques improved the predictive power of the models over standard linear models. However, urban areas were more sensitive than rural ones to the impact of cluster displacement.

Next Steps

Building on this work, further improvements and refinements in input data and methodology can be anticipated by taking advantage of the newly available 1km and finer resolution covariates, integrating DHS survey data with those from other survey and non-survey sources, and extending the analysis approach to evaluate change by using multiple surveys over time. This report has shown the potential of the mapping approach for three countries and four variables. If the approach were to be adopted as a standard method for map production and distribution, planning would be required to implement scaling up the mapping to other countries and variables, and addressing the issues that have been raised in mapping urban areas. This will require longer-term planning methods and organization for implementation, training, regularity of update, covariate selection and update, variables to be mapped, data hosting and storage, as well as guidance on using the new map surfaces and estimating the uncertainty for program planning and decisionmaking.

1. Background and Objectives

Improved understanding of geographic variation and inequity in health status, wealth, and access to resources within countries is increasingly recognized as central to meeting development goals. Development and health indicators assessed at national scale can often conceal important inequities, with the rural poor often the least well represented. As international funding for health and development comes under pressure, the ability to target limited resources to underserved groups becomes crucial. Monitoring demographic, access, and health status inequalities for targeting interventions and measuring progress towards health and development goals require a reliable, detailed evidence base. While high-resolution spatial data on population distributions in resource poor areas are now becoming available (e.g., www.worldpop.org.uk), comprehensive, contemporary, and spatially detailed information on demographic and health attributes of those populations currently remain usable only at aggregated regional levels through national household surveys such as those conducted by The Demographic Health Survey (DHS) Program.

The DHS Program has been a leader in collecting and providing cluster-randomized survey data on core development indicators. In addition to the standard open-source data files in which household and individual survey results can be tabulated by first-order sub-national regions (e.g., at province or state level) and urban/rural strata, more recent surveys now provide geocoded data for individual clusters. The availability of the Global Positioning System (GPS) coordinates for DHS (and Malaria Indicator Survey (MIS) and AIDS Indicator Survey (AIS)) clusters provides highly local scale information that can be linked with survey outputs for quantifying demographic and health status heterogeneities and inequities.

Many recent studies have utilized a range of interpolation approaches to derive continuous estimated surfaces of variables of interest from DHS household surveys; these were reviewed in The DHS Program Spatial Analysis Report (Burgert 2014). The number of studies conducted, as well as interest from donors and governments, indicates a demand for interpolated maps created from survey locations to create the estimates for smaller areas than are currently provided through The DHS Program. How these maps should be produced, whether reliable and accurate maps are feasible in both rural and urban geographies, how data displacement influences anonymity, what form the outputs should take, and which variables should be mapped are all questions that remain to be answered.

This report addresses three overarching aims:

1. To explore the potential of Bayesian model-based geostatistics for the production of interpolated surfaces from GPS cluster located survey data.
2. To explore the impact of DHS cluster displacement on the production of interpolated surfaces.
3. To investigate the potential for novel methodologies and covariates to address the challenge of mapping within urban areas.

1.1 Objective 1: Exploring the Potential of Bayesian Geostatistics for Interpolating DHS Survey Data

Previously, Bayesian geostatistics (see Box 1) have been shown to be valuable methodologies for the production of interpolated surfaces of malaria and poverty prevalence from GPS-located survey data (Gething et al. 2011; Tatem et al. 2014); they therefore represent a good method for extension to the mapping of other survey variables. The approach exploits spatiotemporal relationships within the data, leverages ancillary information from geospatial covariates, and rigorously handles uncertainties at all stages to generate robust output surfaces with accompanying confidence intervals.

Box 1. Bayesian Geostatistics

The term *geostatistics* refers to a collection of statistical tools that have been developed to aid the understanding and modeling of spatial variability, with the principal motivation of predicting unsampled values dispersed in space (also termed *interpolation*). The most widely used tool, *Kriging*, is an interpolation approach whereby predicted values are made at unsampled locations based on a weighted combination of nearby data points. Unlike simpler interpolation algorithms, Kriging provides "optimum" accuracy of predicted values by identifying the most suitable weights for each data point. This, in turn, is achieved by characterizing the degree of correlation between points across space by using a *variogram* function.

Bayesian inference is a method of statistical inference based on *Bayes' theorem*. It allows the combination of prior knowledge (or lack of it) with new information as data are included, and is widely used as a flexible, theoretically rigorous approach to fitting statistical models based on sampled datasets.

Bayesian geostatistics refers to the implementation of geostatistical models with Bayesian methods of inference. Uncertainty in the data (i.e., sampling variation) and in the fitted model parameters (such as the shape of the variogram or autocorrelation function, and relationships with covariates) is inferred and propagated, so that it can be measured in the output predictions. In practical terms, this allows a convenient way of propagating uncertainty through all stages of the model fit, and the representation of this uncertainty in mapped outputs as a *posterior distribution* for each predicted pixel value.

This report presents the analyses of the production of interpolated surfaces through model-based geostatistics by addressing the following objectives:

- Select four indicators to study; these should fit the indicator criteria set forth in DHS Spatial Analysis Report 9 "Spatial Interpolation with Demographic and Health Survey Data: Key considerations" (Burgert 2014).
- Select three pilot countries (one survey year each) to study; these should be geographically diverse but not surveys that include HIV testing to protect the identity of individuals in the non-displaced data in the analysis.
- Select an appropriate list of covariates openly available on an international, regional, or country scale, and conduct statistical tests for their association with the indicators for spatial interpolation.
- Create basic methodological approaches and apply to indicators/countries selected.
- Validate the methodological approach by using proven techniques such as root-mean-square error (RMSE) and coverage probability of prediction interval (e.g., leave-one-out).

1.2 Objective 2: Exploring the Impact of DHS Cluster Displacement on the Production of Interpolated Surfaces

DHS survey data are subject to a standardized geographic masking procedure that ensures data are confidential and that the individual privacy of health information is maintained. This procedure applies a random geographical displacement to the reported latitude and longitude coordinates of each survey cluster location. A key question is whether and how this displacement affects the potential of geostatistical and other geospatial approaches for generating reliable, interpolated surfaces from DHS variables.

This report presents the results of a series of analyses that investigate displacement effects via the following objectives:

- Using non-displaced DHS survey data as a reference point, generate a set of randomly generated displaced datasets by using the standard DHS displacement protocol.
- Explore the effect of displacement on statistical properties of data.
- Explore the effect of displacement on model-based geostatistics (MBG)-derived interpolated surfaces, as generated under Objective 1, by implementing separate MBG models for each iteration of the large set of randomly displaced data, and comparing the performance of each to the reference (non-displaced) set.

1.3 Objective 3: Investigating the Potential for Novel Methodologies and Covariates to Address the Challenge of Mapping within Urban Areas

With over half of the world's population now living in urban areas, and the highest rates of urbanization seen in the lowest income countries and particularly in sub-Saharan Africa, the need for accurate mapping of the demographics and health of urban populations is increasing. Urban areas typically exhibit substantial heterogeneities in health and development indicators. Being able to capture these heterogeneities will substantially improve the relevance and accuracy of output gridded surfaces. However, there are a set of limitations that must be overcome before this can be achieved:

- The spatial resolution and scales of variation captured by the majority of covariate layers remains insufficient for accurate characterization of within-urban variations that relate to differences in demographic and health indicators.
- The displacement of cluster centroids of up to 2km in urban areas represents a substantial shift compared to the spatial scales of variation that exist within urban areas in terms of demographics and health, e.g., within an urban area, rich areas and informal settlements can be neighbors.
- Official government definitions of what is "urban" (generally used to define urban *vs.* rural clusters in DHS survey design) vary from country to country; there can sometimes be substantial differences, which potentially lead to clusters that would be defined as urban in one country being classed as rural in another, and vice-versa; this makes cross-country comparisons difficult.
- In the highly patchy nature of urban landscapes, it may be challenging to meet the assumptions required for traditional linear modeling approaches.

For this objective, we outline exploratory analyses and methods for addressing these issues. With the increasing availability of high resolution, accurate, potential covariate layers that capture within urban variability, the potential exists to undertake mapping of urban populations' demographic and health metrics at higher resolution than outlined in previous sections. Further, the effects of the 2km maximum displacement on mapping accuracies are unknown when combined with such layers, or how smaller buffers might improve mapping. In addition, with the availability of multiple standardized satellite-derived "urban" definition maps, the implications of switching to one that would define urban/rural cluster definitions are unknown in terms of numbers of clusters that would switch classes. Finally, we explore novel analytical approaches to analyzing data from urban areas that may be better able to accommodate the variable urban landscape than the traditional linear modeling approaches.

2. Selection of DHS Indicators and Exemplar Countries

The selection of indicators for interpolation and pilot countries that leveraged the discussions at the consultative meeting in June 2013 are summarized in DHS Spatial Analysis Report 9. These should also include some of the most commonly used DHS indicators in the areas of nutrition, HIV, malaria, and fertility. In addition to these criteria, it was deemed desirable to select indicators from different indicator types (knowledge, behavior, state, or biomarker), levels (individual (child/women/men) or household), and topic areas. The goal was also to identify some variables that were likely to have strong statistical relationships with background environmental factors, as well as others that were less likely to be strongly environmentally mediated. With these parameters, the following indicators were chosen for this study: percentage of women 15-49 who were tested for HIV in the previous 12 months; percentage of children 6-59 months who are stunted (at least 2 standard deviations below the median height for children of the same sex and age according to the WHO standard); percentage of children 6-59 who are anemic (mild, moderate, or severe); and percentage of households with access to improved sanitation according to international standards.

Country and survey year selection focused on four criteria: no HIV testing conducted during the survey, survey conducted since 2008, countries located on the African continent, and countries large enough to exhibit variation in indicators across the country. Recent surveys and surveys in Africa allow for leverage of the Oxford University covariate library, which has a larger selection of more recent data and a strong African focus. These criteria led to the selection of the 2008 Ghana DHS, the 2010 Tanzania DHS, and the 2011 Uganda DHS (ICF International 2008-2011).

3. National-Level Geostatistical Mapping

3.1 Methods

3.1.1 Overview

Building on techniques that were originally conceived for detailed mapping of malaria prevalence (Gething et al. 2011; Gething et al. 2012), a body of theory known as Bayesian model-based geostatistics (MBG) (Diggle and Ribeiro 2007; Diggle, Tawn, and Moyeed 1998) is the basis of the approach used in this project. In an MBG framework, the observed cluster-level variation in the indicator variable of interest is explained by four components.

1. Sampling error, which can often be large given the small sample sizes in individual clusters, is represented with a standard sampling model. This might, for example, be a binomial model where cluster-level data include a number of "positive" households or individuals (such as children with anemia or households with adequate sanitation) from a total number sampled.
2. Some non-sampling variation can often be explained with fixed effects, whereby a multivariate regression relationship is defined by linking the indicator variable with a suite of geospatial covariates.
3. Additional non-sampling errors not explained by the fixed effects are usually spatially autocorrelated; this is represented with a random effects component. A spatial multi-variate normal distribution known as a Gaussian Process is employed, and parameterized by a spatial covariance function.
4. Finally, any remaining variation not captured by these components is represented with a simple Gaussian noise term equivalent to that employed in a standard non-spatial linear model.

The full model output includes for every pixel on the mapped surface, a posterior distribution for the predicted variable that represents a complete model of the uncertainty around the estimated value. These can be summarized with a point estimate (such as the posterior mean) to generate a mapped surface. Additional summary statistics from each posterior distribution can also be mapped to create maps of the posterior standard deviation or 95% credible intervals that demonstrate geographical areas with more or less uncertainty.

This section will describe the preparation of point-georeferenced data on each selected DHS indicator, the assembly and exploration of a suite of gridded geospatial covariate layers, and the use of these inputs in a series of bespoke MBG models to generate final maps for each indicator.

3.1.2 Preparation of Georeferenced DHS Indicator Data

In most DHS household surveys, the sampling clusters are the primary sampling unit (PSU), which are preexisting geographic areas known as census enumeration areas (EAs). The boundaries of the EAs are defined by the country's census bureau, as are the urban and rural status of each cluster. An EA can be a city block or apartment building in urban areas, while in rural areas it is typically a village or group of villages. The population and size of sampled clusters vary between and within countries. Typically, clusters contain 100-300 households, of which 20-30 households are randomly selected for survey participation. The estimated center of each cluster is recorded as a latitude/longitude coordinate, obtained from a GPS receiver or derived from public online maps or gazetteers. The actual physical size or boundaries of the survey cluster are not usually known, although in recent years it has become more common for countries to have census EA boundary files that are then used to calculate the center of the EA.

The georeferenced datasets can be linked to individual and household records in DHS household surveys through unique cluster identifiers. To insure confidentiality, the geographic coordinates of the cluster are randomly displaced prior to dataset release (Burgert et al. 2013).

The publicly available survey data and displaced cluster GPS coordinates for the Ghana 2008 DHS, Tanzania 2010 DHS and Uganda 2011 DHS were obtained. The following sections describe the processing steps used to construct the cluster level variables of interest that would be mapped.

3.1.2.1 Access to HIV testing in women

In the individual women's questionnaire, women 15-49 years are asked if they had ever been tested for HIV and if they had been tested within the last 12 months. This indicator is not expected to be linked to the biophysical environment, but may have structured geographic variation across a country since access to testing may vary from one place to another.

Numerator: Number of women aged 15-49 who accessed HIV testing in the last 12 months.
Denominator: Number of women aged 15-49.

3.1.2.2 Stunting in children

Stunting is defined as children with a height at least two standard deviations (SD) below the median height for children of the same sex and age in an internationally standard index defined in the WHO Child Growth Standards (2006). Stunting is a measure of chronic malnutrition, and in some countries may be environmentally linked where the combination of poverty and low agricultural productivity of local land limit calorific intake in children.

Numerator: Number of children aged 6-59 months who are stunted (<2 SD).
Denominator: Number of children aged 6-59 months.

3.1.2.3 Anemia prevalence in children

Anemia is characterized by a low level of hemoglobin in the blood and is routinely diagnosed with a blood test. The DHS Program tests women (15-49 years) and children (usually 6 months to 5 years) for anemia through a finger prick or, in the case of young children, heel prick blood testing with the HemoCue blood hemoglobin testing system. Anemia is considered an underlying cause of many poor health outcomes such as malnutrition or parasite infection (such as malaria), and in many countries is environmentally linked. Hemoglobin levels in individuals vary naturally due to differences in elevation, i.e., available oxygen concentrations in the air. This is taken into account in DHS survey data to provide elevation-standardized, geographically comparable indices.

Numerator: Number of children aged 6-59 months who are anemic ("anemia level" recorded as mild, moderate, or severe).
Denominator: Number of children aged 6-59 months.

3.1.2.4 Access to improved sanitation

A household's access to improved sanitation is an important development indicator. Improved sanitation is defined in this study as a toilet facility that separates the waste from human contact and is used only by members of the household (i.e., not shared). Toilet facilities that qualify as improved include flush/pour flush to piped sewer system/septic tank/pit latrine, ventilated improved pit latrine, and pit latrine with slab.

Numerator: Number of households with improved sanitation.
Denominator: Number of households.

Preparation of geospatial covariate layers

As described in Section 3.1, an important aspect of geostatistical modeling is the exploitation of geospatial covariates that are correlated with the outcome of interest, and can partially explain variation in that response and allow for more accurate predictions across the map. A suite of covariates were chosen from existing libraries held at the University of Oxford, based on factors that have previously been shown to correlate with demographic and health indicators in different settings.

3.1.2.5 Defining a set of candidate covariate layers from existing geospatial products

A range of covariate grids were included for testing as possible explanatory covariates across all four DHS test indicators. The features, sources, and naming conventions for each are outlined in Table 1, with further details provided below.

Table 1. Details of geospatial covariates assembled for use in mapping

Dataset Name	Dataset Description	Continuous or Categorical	Original Data Source	Date
access	Travel time to cities with > 50k via all transport methods	Continuous	http://forobs.jrc.ec.europa.eu/products/gam/	2000
afripop	WorldPop (AfriPop) population density dataset	Continuous	www.worldpop.org.uk	2010
grump	GRUMP population density	Continuous	http://sedac.ciesin.columbia.edu/data/collection/grump-v1	2000
GPW	GPW population density	Continuous	http://sedac.ciesin.columbia.edu/data/set/gpw-v3-population-count/data-download	2010
aridity	Mean annual aridity	Continuous	http://csi.cgiar.org/Aridity/	1950-2000
pet	Mean annual Potential Evapotranspiration	Continuous	http://csi.cgiar.org/Aridity/	1950-2000
Lights.2010	Global DMSP-OLS Nighttime Lights Time Series	Continuous	http://ngdc.noaa.gov/eog/	2010
Lights.2012	VIIRS Nighttime Lights-2012	Continuous	http://www.ngdc.noaa.gov/dmsp/download.html	2012
elevation	Shuttle Radar Topography Mission (SRTM) Near-global Digital Elevation Models (DEMs)	Continuous	http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10008_1	2000
evi	Enhanced vegetation index	Continuous	http://modis.gsfc.nasa.gov/	2001-2014
Lst.day	Land surface temperature in the daytime	Continuous	http://modis.gsfc.nasa.gov/	2001-2014
Lst.night	Land surface temperature in the nighttime	Continuous	http://modis.gsfc.nasa.gov/	2001-2014
midir	Middle Infrared reflectance	Continuous	http://modis.gsfc.nasa.gov/	2001-2014
TSI	Temperature Suitability Index	Continuous	http://modis.gsfc.nasa.gov/	
TCB	Tasseled-cap brightness	Continuous	http://modis.gsfc.nasa.gov/	2001-2014
TCW	Tasseled-cap wetness	Continuous	http://modis.gsfc.nasa.gov/	2001-2014
precip	Average monthly rainfall	Continuous	http://www.worldclim.org/	1950-2000

Travel times

The EU Joint Research Center maintains a gridded surface that estimates accessibility, measured in likely travel times (via all transport methods), to cities with more than 50,000 inhabitants. In practice, this provides a useful composite measure of the extent to which regions are rural vs. urban as well as the degree of their connectedness to the national system of transportation. Areas near major roads, for example, would be relatively well connected, even if they were some distance from major cities. More details of this geospatial layer can be found at http://forobs.jrc.ec.europa.eu/products/gam/.

Population density

Gridded data on population density across the three pilot countries were constructed from satellite-derived settlement maps and available census data as part of the WorldPop project: (www.worldpop.org.uk). Alternative population grids, from the Global Rural Urban Mapping Project v1 (GRUMP, http://sedac.ciesin.columbia.edu/data/collection/grump-v1http://sedac.ciesin.columbia.edu/data/set/grump-v1-population-density) and Gridded Population of the World v3 (GPW, http://sedac.ciesin.columbia.edu/data/collection/gpw-v3) were also investigated.

Aridity and Potential Evapotranspiration (PET)

The CGIAR Consortium maintains high-resolution global raster climate data related to evapo-transpiration processes and rainfall deficit for potential vegetative growth. These are based on data from the WorldClim project and ultimately from weather station data that has been interpolated by using covariates such as altitude. These grids, extracted for the three pilot countries, allowed for differentiation of areas with adequate rainfall and moisture regimes to sustain agriculture vs. those where drier and more arid conditions prevail (http://csi.cgiar.org/Aridity/).

Nightlights

Defense Meteorological Satellite Program Operational Linescan System (DMSP OLS) annual composite satellite nightlight data for 2010 were obtained for the three pilot countries: http://www.ngdc.noaa.gov/dmsp/sensors/ols.html. Also Visible Infrared Imaging Radiometer Suite (VIIRS) composite nightlight imagery for 2012 was obtained (http://ngdc.noaa.gov/eog/). These surfaces allow differentiation of regions based on both the density of population and the degree of electrification of dwellings, commercial and industrial premises, and infrastructure.

Elevation

A digital elevation model (DEM) derived from the NASA Shuttle Radar Topography Mission (SRTM) Near-global Digital Elevation Models (DEMs) was obtained for the three pilot countries, differentiating high from low altitude regions (http://webmap.ornl.gov/wcsdown/wcsdown.jsp?dg_id=10008_1).

MODIS Climatic/environmental conditions

NASA's Moderate Resolution Imaging Spectroradiometer (MODIS) generates high-resolution satellite imagery on four key measures on environmental conditions. The measures that were extracted for the three pilot countries: land surface temperature (LST), the enhanced vegetation index (EVI) and 'Tasseled cap' brightness and wetness products (http://modis.gsfc.nasa.gov/).

Precipitation

The WORLDCLIM project (http://www.worldclim.org/) has generated global surfaces of total monthly precipitation at a notional 1km resolution by interpolating long-term records from a network of ground-based meteorological stations, in conjunction with digital elevation model data.

Temperature Suitability Index

This is a relative index derived from MODIS land-surface temperature data that indicates the degree of suitability of annual temperature regimes in each pixel for sustaining vector-borne disease transmission, with a particular focus on *P. falciparum* malaria. The variable takes into account both overall temperatures and the seasonal patterns of fluctuating temperatures.

3.1.2.6 Defining and implementing a standardized grid format

The geospatial data sources described above were obtained in a variety of spatial resolutions and geographic extents. In addition, the land-sea templates inevitably differed slightly between products, so that the precise definition of coastlines and the inclusion or exclusion of small islands and peninsulas was not consistent. These factors precluded the direct use of these data in a single spatial model. To overcome these incompatibilities and to generate a fully standardized suite of input grids on an identically defined geographic template, a processing chain was developed with the following stages. First, each input data source was re-projected, where necessary, by using a standardized equirectangular Plate Carrée projection under the World Geodetic System 1984 coordinate system. Second, input grids that were defined at differing spatial resolutions were re-sampled to 5×5 km. Third, grids were either extended or clipped to match a standardized extent. Finally, a bespoke algorithm was developed that compared each rectified and re-sampled grid to a "master" land-sea template for the three pilot countries. This used a simple interpolation and/or clipping procedure to align new grids to this master template, which ensured that all coastline was perfectly consistent on a pixel-by-pixel basis.

3.1.3 Exploratory Analysis

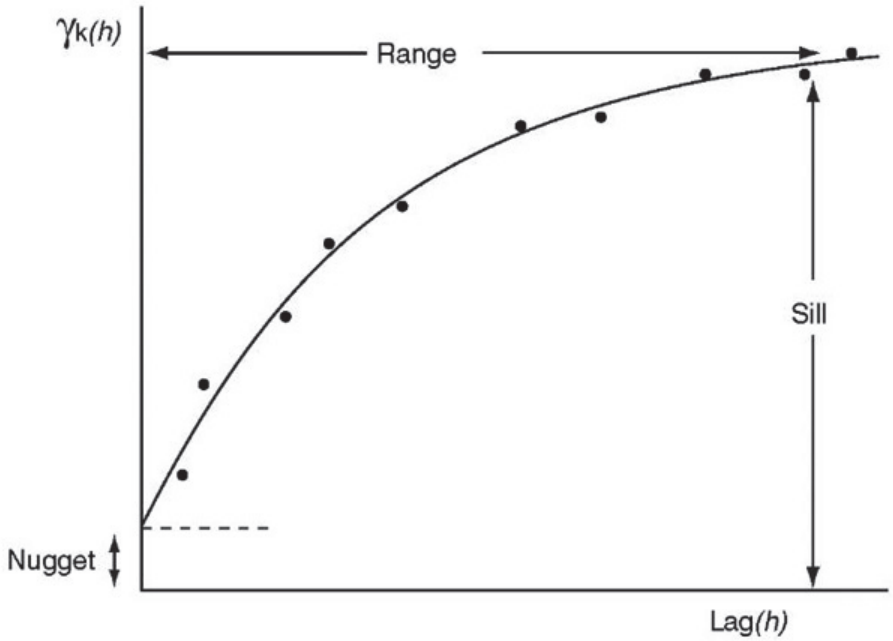
3.1.3.1 Histograms and variography

Two basic exploratory analyses were undertaken for each DHS indicator and each country. First, simple empirical histograms were generated to assess the statistical distribution which can be useful in planning the optimal structure of the geostatistical model, e.g., whether a standard binomial sampling model is appropriate, or whether variations such as a zero-inflated binomial would be more appropriate.

Second, the spatial autocorrelation structure in each outcome variable was assessed via an empirical variogram. An example of a variogram is shown in Figure 1. This function plots semivariance, $\gamma(h)$ (the average dissimilarity between two data points, for example two DHS clusters) against spatial lag, h, (the geographical distance separating two points). Where data are spatially structured, a characteristic variogram form is for semivariance, $\gamma(h)$, to steadily increase with increasing lag, h, as shown in Figure 1. Conversely, data with no spatial structure lead to a flat variogram. This is intuitive since it demonstrates that as points with progressively larger separation distances are compared, their level of dissimilarity increases. In other words, observations closer together are likely to be more similar in value than those further apart. This pattern tends to reach a limit at a certain level of semivariance (the "sill", Figure 1) beyond which additional spatial separation has no further effect on dissimilarity. This is the "range" of spatial autocorrelation (Figure 1), beyond which points are considered spatially independent. In practice, spatial variables tend to display both spatially structured and random variation. The value of semivariance at the intercept (termed the "nugget", Figure 1) is indicative of the amount of non-structured, random variation displayed by the

variable (i.e., variation that persists even over the shortest separation distances). In simple terms, the smaller this nugget value relative to the plateau of the semivariogram, the larger the fraction of variation that is spatially structured. Where the ultimate objective is to use a spatial interpolation technique as part of a predictive model, larger fractions of spatially structured variation will tend to translate into better predictive precision, since more of the variation is predictable rather than apparently random.

Figure 1. Example of a typical variogram showing the different features as described in the text



3.1.3.2 Covariate selection

The covariate assembly and generation described in section 3.3.1 yielded a starting set of 17 candidate geospatial covariates for inclusion in the multivariate fixed-effect component of the geostatistical model for each indicator. Although the DHS surveys collect a large number of variables from each survey cluster that can act as covariates in predictive models, such data are not useful for spatial interpolation because observed values of these variables are not available across the entire mapping region. As with standard (i.e., non-spatial) regression models, selecting the optimal set of covariates is vital to maximize the ultimate predictive accuracy of the model. Including too few informative covariates results in loss of exploratory power, while the inclusion of too many causes the resulting high-dimensional multivariate model to over-fit the data. This explains noise rather than signal and, ultimately, reduces predictive accuracy. Because full geostatistical models are extremely time-consuming to fit, the standard practice is using simpler non-spatial models to determine the optimum covariate selection for subsequent inclusion in the full spatial modeling framework. Techniques such as stepwise variable selection are commonly used, where a candidate set of covariate sets is built up by progressively adding new candidate covariates to a model (forward selection) or subtracting them from an initial inclusive set (backwards selection), and then deciding to keep or discard each new covariate based on its impact on model fit. However, these techniques are known to be sensitive to the order in which variables are added or removed, and therefore risk the generation of arbitrary final selections. For this analysis, a more exhaustive approach was developed that explored a much wider proportion of all possible combinations before identifying the optimum covariate sets. This proceeded in the flowing six steps, which were repeated for each indicator/country:

1. Two new versions of each of the original 17 covariates were generated: a square and square-root transform yielding 17×3 = 51 candidate covariates. This was done to explore whether these non-linear transforms of each original covariate may be more strongly correlated with the indicators than the original.
2. For each of the 17 covariate types, the best performing version (original, square, square-root) was identified by building a simple linear model for each indicator and assessing the predictive R-squared statistic (PR^2). Like its better known counterpart R^2, PR^2 approximates the proportion of variation in the response variable that is explained by the uni- or multivariate model. Importantly, however, PR^2 is calculated out-of-sample by assessing the model's ability to predict unobserved data rather than those used to fit the model itself. As such, it can detect when models begin to over-fit. This stage reduced the 51 covariate-transform combinations down to a set of 17, which consisted of a variety of squared, square-rooted, and untransformed covariates.
3. Very often, individual covariates are more informative when allowed to interact. As a hypothetical example, levels of vegetation greenness associated with fertile land may be informative of stunting - via effects on nutritional status - but only in rural settings. In this scenario, the interaction between a satellite-derived vegetation index and a measure of population density could tell us more about stunting rates than the vegetation index alone. To explore this, every possible pair-wise interaction (simply the multiple of each covariate with every other) was generated. This yielded 272 additional layers to add to the 17 non-interaction layers, for a total of 289.
4. Having defined this expanded set of 289 possible input covariates, the next objective was to assess which combination of these yielded the best performing multivariate model. Since there are millions of potential combinations of 289 variables, assessing each combination would be computationally prohibitive. To make this problem tractable, the univariate PR^2 of all 298 input covariates were assessed individually, and the top 20 retained.
5. Multivariate models were defined and tested (using PR^2) for every possible combination of these top 20 covariates, with no restriction on the final size of the model. This meant that all possible models with 1 to 20 covariates were assessed, for a total of 1,048,575[1] different models.
6. Finally, the best multivariate model was identified from the set of 1,048,575 based on the largest PR^2, and that covariate suite was put selected for inclusion in the full geostatistical model.

3.1.4 Geostatistical modeling

3.1.4.1 Model structure

The initial model-based geostatistical structure is a class of generalized linear mixed model, with an approximation of a multivariate normal random field (i.e., a Gaussian Process) used as a spatially autocorrelated random effect term (Diggle and Ribeiro 2007; Diggle, Tawn, and Moyeed 1998).

The indicator rate (proportion of 'positive' individuals or households according to the particular indicator of interest) $Y(x_i)$, at each location in the pilot country x_i was modeled as a transformation $g(\cdot)$ of a spatially structured field superimposed with additional random variation $\epsilon(x_i)$. The count of positive individuals or households N_i^+ from the total sample of N_i in each survey cluster was modeled as a conditionally independent binomial variate given the unobserved underlying $Y(x_i)$ value. The spatial component was represented by a stationary Gaussian process $f(x_i, t_i)$ with mean μ and covariance C. The unstructured

[1] Calculated by taking ($2^{20} - 1$)

component $\epsilon(x_i)$ was represented as Gaussian with zero mean and variance V. Both the inference and prediction stages were coded with the Integrated Nested Laplace Approximations (INLA) framework, primarily in R (Blangiardo et al. 2013; Fong, Rue, and Wakefield 2009; Rue, Martino, and Chopin 2009).

Mean and covariance definition

The mean component μ was modeled as a linear function of the *n* geospatial covariates identified in the preceding stage $\mu = \beta\mathbf{X}$, where $\mathbf{X} = 1, X_1(x), ..., X_n(x)$ was a vector with a constant and the covariates indexed by spatial location x,and $\beta = \beta_0, \beta_1, ..., \beta_n$ was a corresponding vector of regression coefficients. Each covariate was converted to *z*-scores before analysis. Covariance between spatial locations was modeled using a Matern covariance function C:

$$C\left(d(x_i;x_j)\right) = \sigma^2 \frac{1}{\Gamma(\upsilon)2^{\upsilon-1}}\left(\sqrt{2\upsilon}\frac{d(x_i;x_j)}{\rho}\right)^{\upsilon} K_{\upsilon}\left(\sqrt{2\upsilon}\frac{d(x_i;x_j)}{\rho}\right)$$

Where $d(x_i;x_j)$ is the geographical separation between two points; σ, υ, ρ are parameters of the covariance function defining, respectively, its amplitude, degree of differentiability, and scale; K_{υ} is the modified Bessel function of the second kind of order υ, and Γ is the gamma function (Antosiewicz 1964; Davis 1964).

3.1.4.2 Model implementation and output

Bayesian inference was implemented using the INLA algorithm to generate approximations of the marginal posterior distributions of the outcome variable $Y(x_i)$ at each location on a regular 5 × 5 km spatial grid across each pilot country and of the unobserved parameters of the mean, covariance function and Gaussian random noise component. At each location, the posterior distribution was summarized with the posterior mean as a point estimate, and the posterior inter-quartile range as a measure of model precision. Maps were generated of each metric in ArcGIS 10.2.

3.1.4.3 Validation

The predictive performance of each model was assessed via out-of-sample validation. We implemented a four-fold hold-out procedure in which 25% of the data points were randomly withdrawn from the dataset, the model run in full with the remaining 75% of data, and the predicted values at the locations of the hold-out data compared to their observed values. This was repeated four times without replacement so that every data point was held out once across the four validation runs. Standard validation statistics were computed as measures of model precision (mean absolute error, MAE) and bias (mean square error, MSE). The MAE quantifies model precision, while the MSE indicates the bias of the model, with values close to zero indicating that the model is unbiased.

3.2 Results

This section presents results for the 12 country-indicators addressed in this project. We organize the presentation of results by indicator, in the following order: access to HIV testing in women; stunting in children; anemia in children; and access to improved sanitation. For each indicator, we first present the results of the **exploratory analysis -** histograms for each country that show the basic statistical distribution of the indicator at cluster level, and variograms that summarize its spatial autocorrelation structure. We also

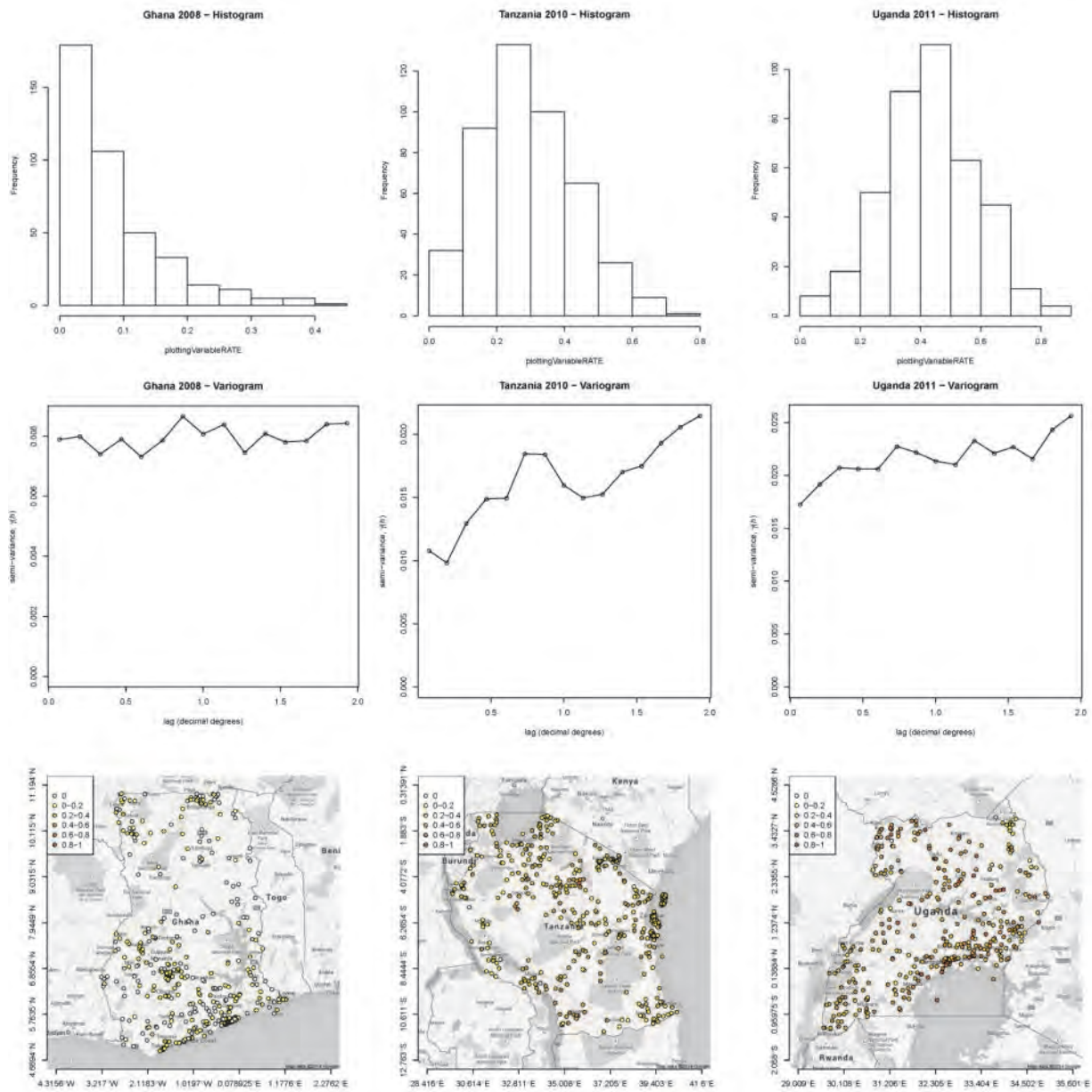
show simple maps of the survey clusters and the indicator value at each cluster. The rationale and interpretation of these types of exploratory plots is described in the methods section. For each indicator, we then present the results of the **covariate selection** exercise, detailing which covariates were selected as the optimum performing set for the given indicator for each country, along with the predictive R^2 achieved. Next, the results of the out-of-sample validation exercise are presented, detailing the MAE and MSE for the given indicator for each country. Finally, each indicator section includes the output maps for each country. As described earlier, the full model output for each pixel is a predictive posterior distribution that describes the level of likelihood associated with all possible values of the indicator variable. As is common practice in Bayesian geostatistics, we use the mean of each posterior as the "point estimate" and summary of the central tendency value of the variable at each pixel; together these form the main predictive map. In addition, we present an accompanying uncertainty map. This summarizes the level of certainty associated with the values shown in the point estimate map by displaying the width of the 95% credible interval for each pixel. Since all indicators in the study are prevalence or proportion variables, all lie on a scale between 0 and 1 (or 0% and 100%). In a situation with complete uncertainty about a pixel's value, the 95% credible intervals would span the entire range – i.e., the true value could lie anywhere between zero and one. Conversely, where a variable is predicted with very high certainty, the width of the 95% credible interval might be very narrow. In other words, there is a 95% probability that the true value lies within a narrow range of possible values.

3.2.1 Access to HIV testing in women

3.2.1.1 Exploratory analysis

Figure 2 shows that for Ghana, there is little variation in values, with the vast majority of clusters showing low proportions of women accessing testing, and virtually no spatial structure evident. Both Tanzania and Uganda show a wider range of proportions, higher rates of access to testing, and some strong spatial dependence evident from the variograms.

Figure 2. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of women accessing HIV testing in the 12 months preceding the survey.



3.2.1.2 Covariate selection

Table 2 summarizes the results of the covariate selection procedure for the indicator on proportion of women accessing HIV testing in the 12 months preceding the survey. Shown for each country are details of the best-performing model, as defined by the exhaustive out-of-sample comparison of all individual and interaction terms across 17 initial candidate covariates. The covariate naming conventions and descriptions can be found in Table 1.

Table 2. Summary output from the covariate selection procedure for the indicator on proportion of women accessing HIV testing in the 12 months preceding the survey. P*R*² is the predictive R-squared statistic. 'sqrt' denotes a square-root transform was applied to the covariate term

	Ghana	Tanzania	Uganda
No. of Covariates	3	6	5
PR²	0.071	0.150	0.142
Covariates in best model	access access **x** GPW lights.2012 **x** precip	TSI access **x** TCB access **x** precip aridity **x**TSI PET **x**TSI TCW **x**TSI	access **x** PET access **x** LST aridity **x** PET aridity **x** LST_night lights.2010 **x** PET

3.2.1.3 Geostatistical model: Validation statistics

Table 3 presents results of the validation exercise for the geostatistical model predicting the indicator on the proportion of women accessing HIV testing in the 12 months preceding the survey. As can be seen, the model for Ghana demonstrated the best performance since it was the least biased (smallest MSE) and with the smallest error magnitude (MAE). All models performed reasonably well with small bias values and average errors around 10 percentage points.

Table 3. Summary validation results from the model-based geostatistical map for the indicator on proportion of women accessing HIV testing in the 12 months preceding the survey. These are outputs from the four-fold out-of-sample validation exercise for each pilot country. MAE = mean absolute error, measuring the precision of predicted values; MSE = mean square error, measuring the bias of predicted values. Both metrics are in the same units as the variables themselves (which are all measured as proportions between zero and one).

	MAE	MSE
Ghana	0.065	0.007
Tanzania	0.104	0.018
Uganda	0.112	0.020

3.2.1.4 Mapped surfaces

Figure 3-5 show the mapped model outputs for access to HIV testing for Ghana, Tanzania, and Uganda, respectively. These results highlight the challenges in mapping a variable with environmental factors that is likely to be largely driven by non-environmental factors. The predictive R^2 values in Table 2 are low, especially for Ghana, for which the resultant prediction map in Figure 3 shows little variation. However, this does reflect the input cluster-level data, and low MAE and MSE values show that the model is precise and unbiased. Table 2 shows that the access covariate was consistently selected, as reflected in the road network patterns in the output maps. For all countries, the maps are predicted with relatively high certainty. The uncertainty maps are predominately blue – with a narrow width of the 95% credible intervals.

Figure 3. (left) Predicted map of indicator on proportion of women accessing HIV testing the 12 months preceding the survey in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.

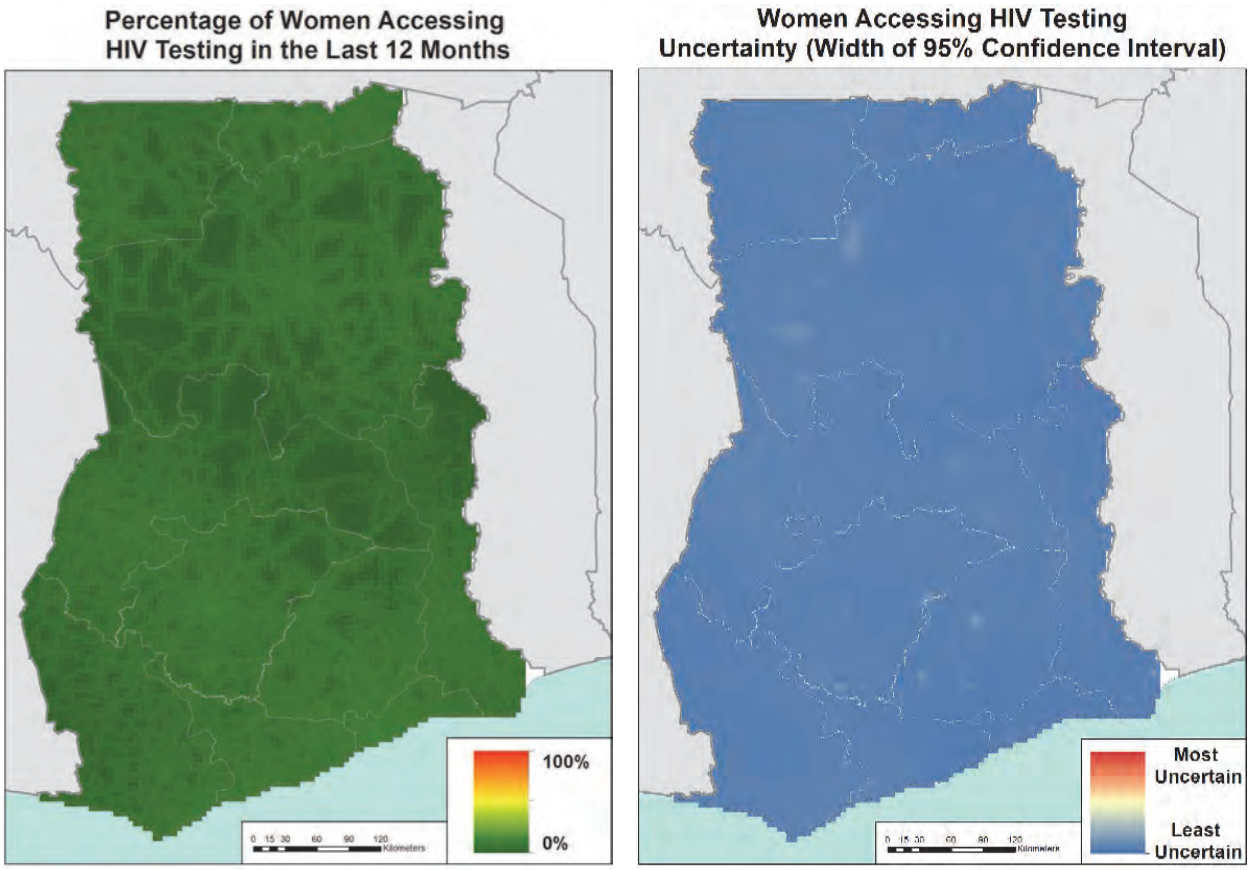


Figure 4. (left) Predicted map of indicator on proportion of women accessing HIV testing the 12 months preceding the survey in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval.

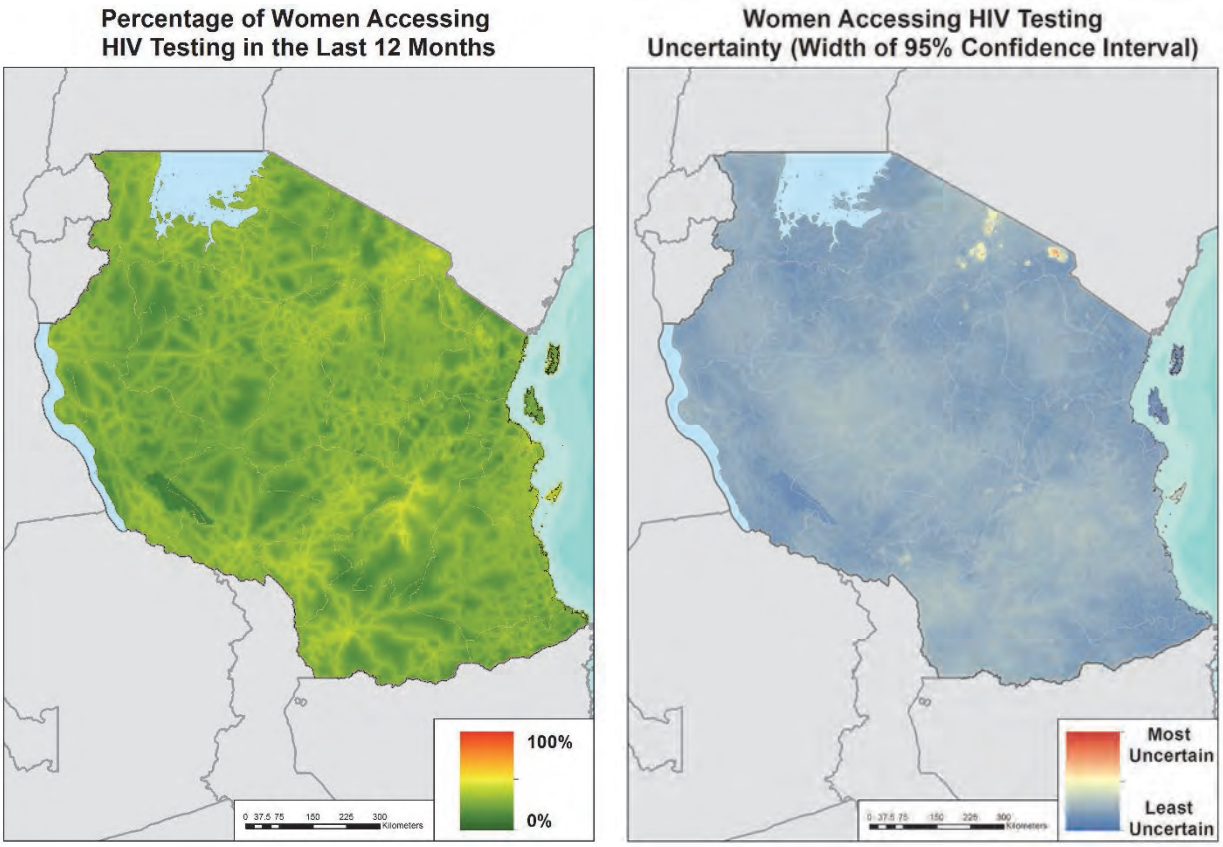
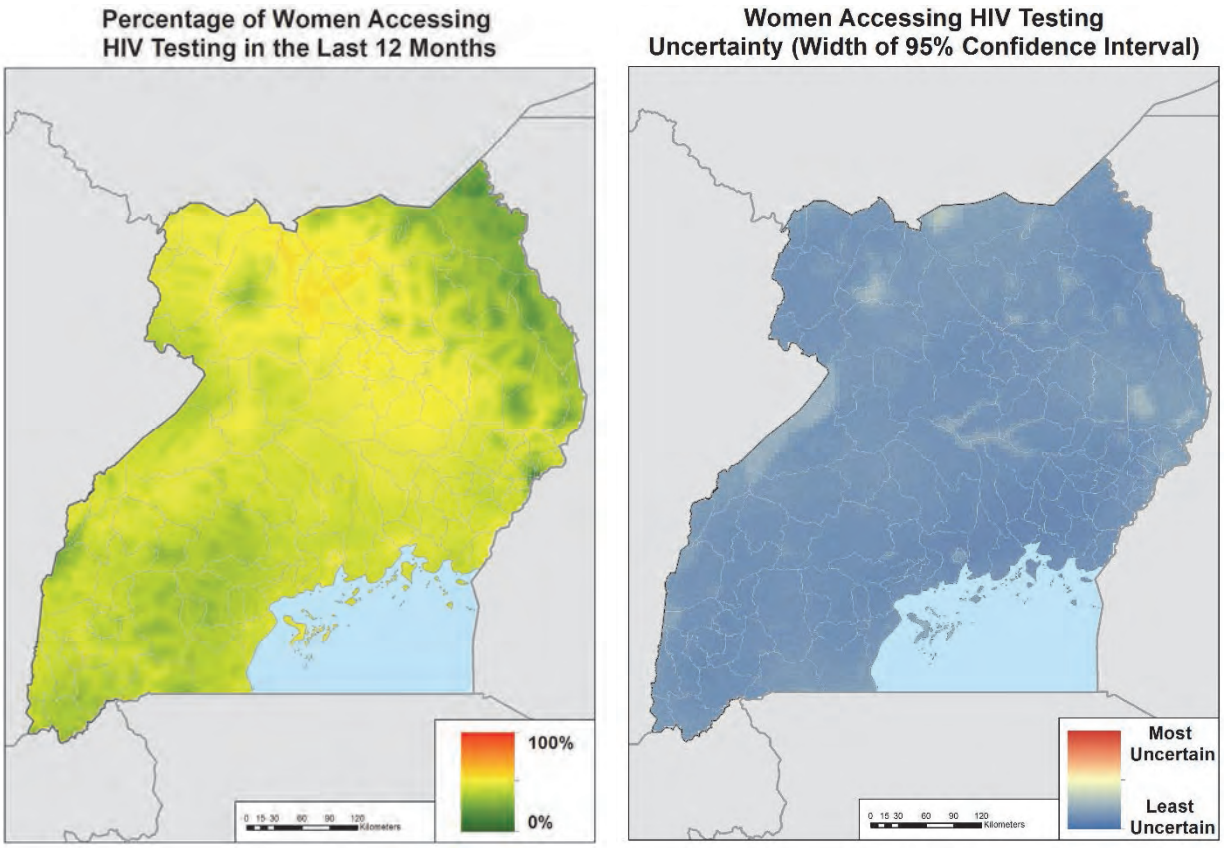


Figure 5. (left) Predicted map of indicator on proportion of women accessing HIV testing the 12 months preceding the survey in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval.



3.2.2 Stunting in children

3.2.2.1 Exploratory analysis

Figure 6. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of children 6-59 months who are stunted.

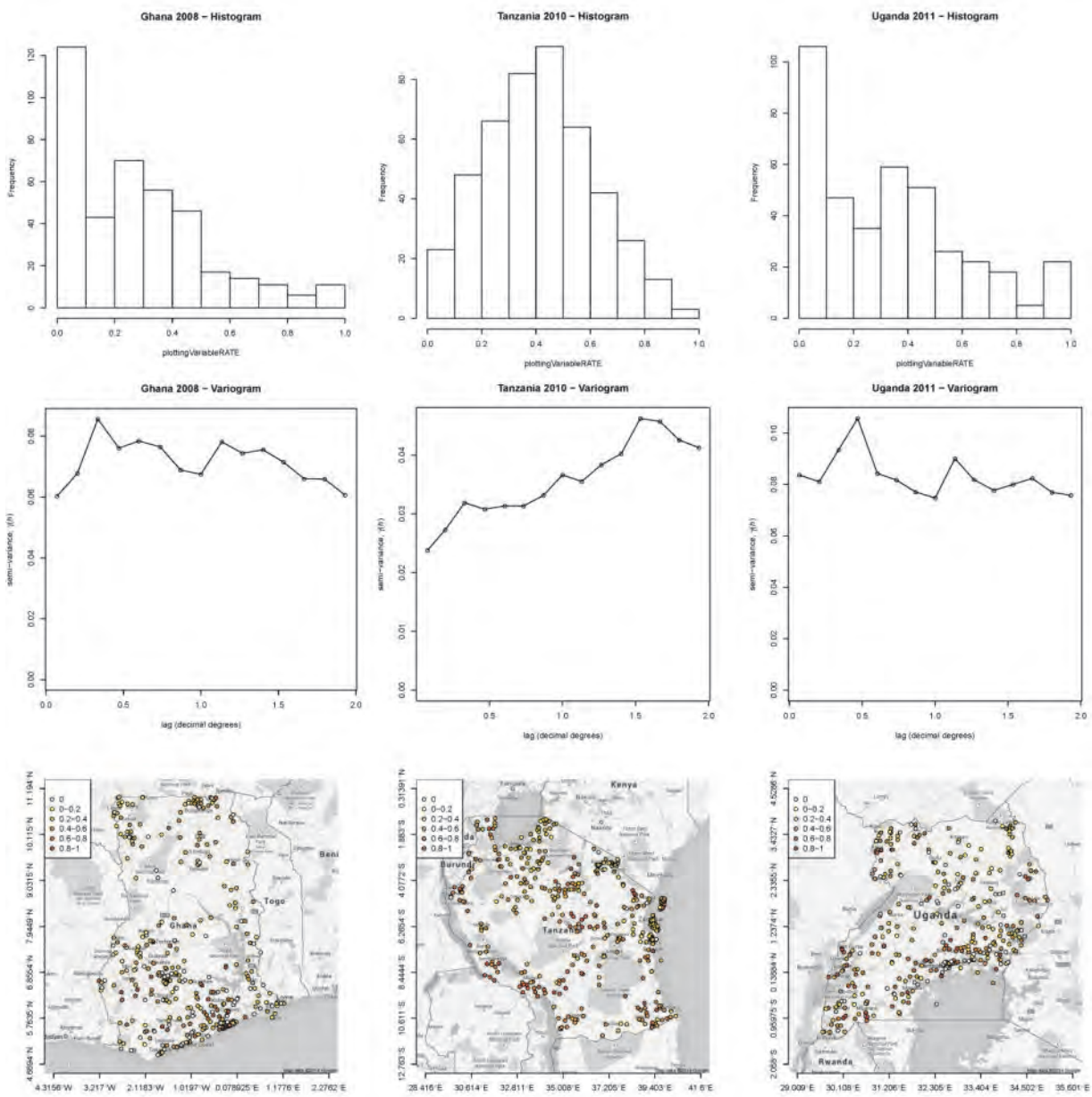


Figure 6 highlights the fact that Tanzania displays a wide range of proportions, with some spatial dependence evident from the variogram, and clear regions of high and low proportions from the map. Ghana and Uganda show very similar distributions of survey outputs, with less clear evidence for spatial dependence and a large number of surveys locations that show no children with stunting.

3.2.2.2 Covariate selection

Table 4 summarizes the results of the covariate selection procedure for the indicator on proportion of children <5 years who are stunted. Shown for each country are details of the best-performing model, as defined by the exhaustive out-of-sample comparison of all individual and interaction terms across 17 initial candidate covariates. The covariate naming conventions and descriptions can be found in Table 1.

Table 4. Summary output from the covariate selection procedure for the indicator on proportion of children <5 years who are stunted. PR^2 is the predictive R-squared statistic. 'sqrt' denotes a square-root transform was applied to the covariate term.

	Ghana	Tanzania	Uganda
No. of Covariates	3	13	9
PR^2	0.068	0.225	0.162
Covariates in best model	(lights.2010)^2 **x** sqrt(LST_night) sqrt(EVI) **x** (TCB)^2 sqrt(EVI) **x** sqrt(precip)	sqrt(GPW) sqrt(grump.2010) sqrt(access) **x** LST sqrt(GPW) **x** sqrt(grump.2010) sqrt(GPW) **x** sqrt(PET) sqrt(GPW) **x** sqrt(LST_day) sqrt(GPW) **x** sqrt(TCB) sqrt(GPW) **x** (TSI)^2 sqrt(grump.2010) **x** sqrt(LST_day) sqrt(grump.2010) **x** sqrt(TCB) sqrt(grump.2010) **x** (precip)^2 lights.2010 **x** (LST_night)^2 lights.2010 **x** (TSI)^2	sqrt(access) sqrt(lights.2010) sqrt(access) **x** sqrt(PET) (aridity)^2 **x** sqrt(GPW) (aridity)^2 **x** sqrt(grump.2010) (aridity)^2 **x** LST_night (elevation)^2 **x** sqrt(lights.2010) sqrt(GPW) **x** (precip)^2 sqrt(lights.2010) **x** LST_night

3.2.2.3 Geostatistical model: Validation statistics

shows validation statistics for the geostatistical model on proportion of children 6-59 months who are stunted. The model for Tanzania displayed the best performance for both bias (MSE) and precision (MAE).

Table 5. Summary validation results from the model-based geostatistical map for the indicator on proportion of children 6-59 months who are stunted. These are outputs from the four-fold out-of-sample validation exercise for each pilot country. MAE = mean absolute error, measuring the precision of predicted values; MSE = mean square error, measuring the bias of predicted values. Both metrics are in the same units as the variables themselves (which are all measured as proportions between zero and one).

	MAE	MSE
Ghana	0.188	0.059
Tanzania	0.111	0.025
Uganda	0.205	0.067

3.2.2.4 Mapped surfaces

Table 4 shows improved predictive R^2 values for Tanzania and Uganda over those for access to HIV testing, but a low value for Ghana. The lights and population covariates are consistently selected in Table 4, driven by the strong urban/rural divides in the cluster-level outcomes. These urban-rural divisions are consequently visible in the outcome predictive maps in Figure 7-9, with substantially lower predicted prevalence of stunting (shown as the green areas) corresponding to the major urban centers of Accra and Kumasi. Uncertainty varied substantially among the three countries, with Ghana being generally least uncertain, and a more heterogeneous level of uncertainty shown in both Tanzania and Uganda.

Figure 7. (left) Predicted map of indicator on proportion of children 6-59 months who are stunted in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.

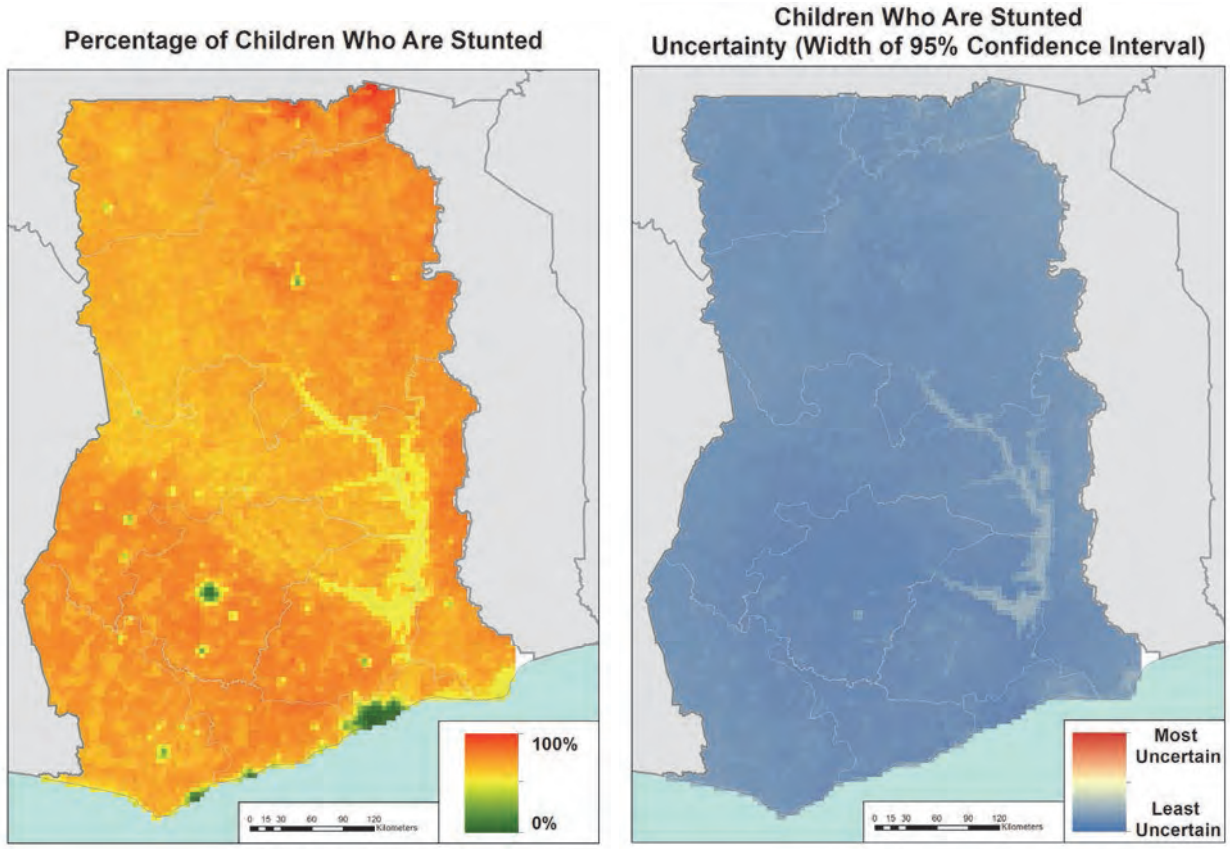


Figure 8. (left) Predicted map of indicator on proportion of children 6-59 months who are stunted in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.

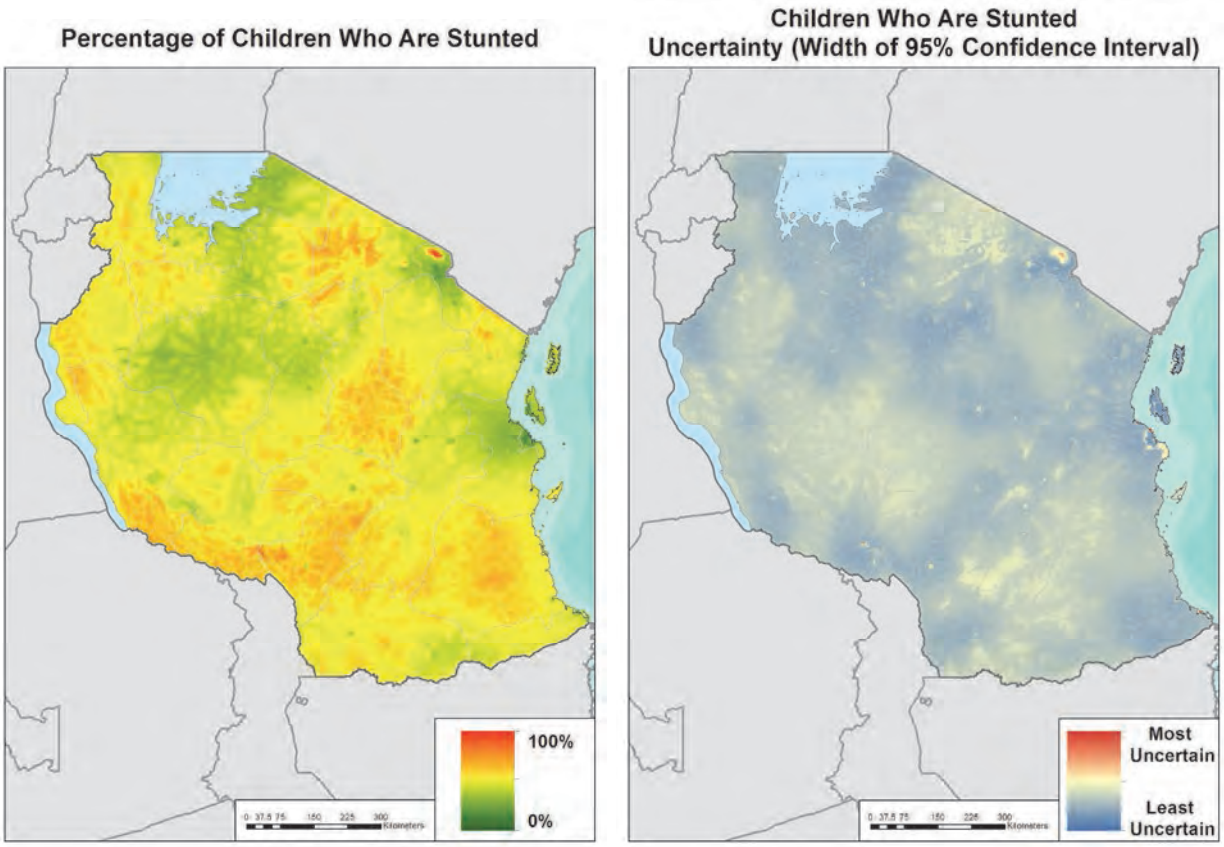
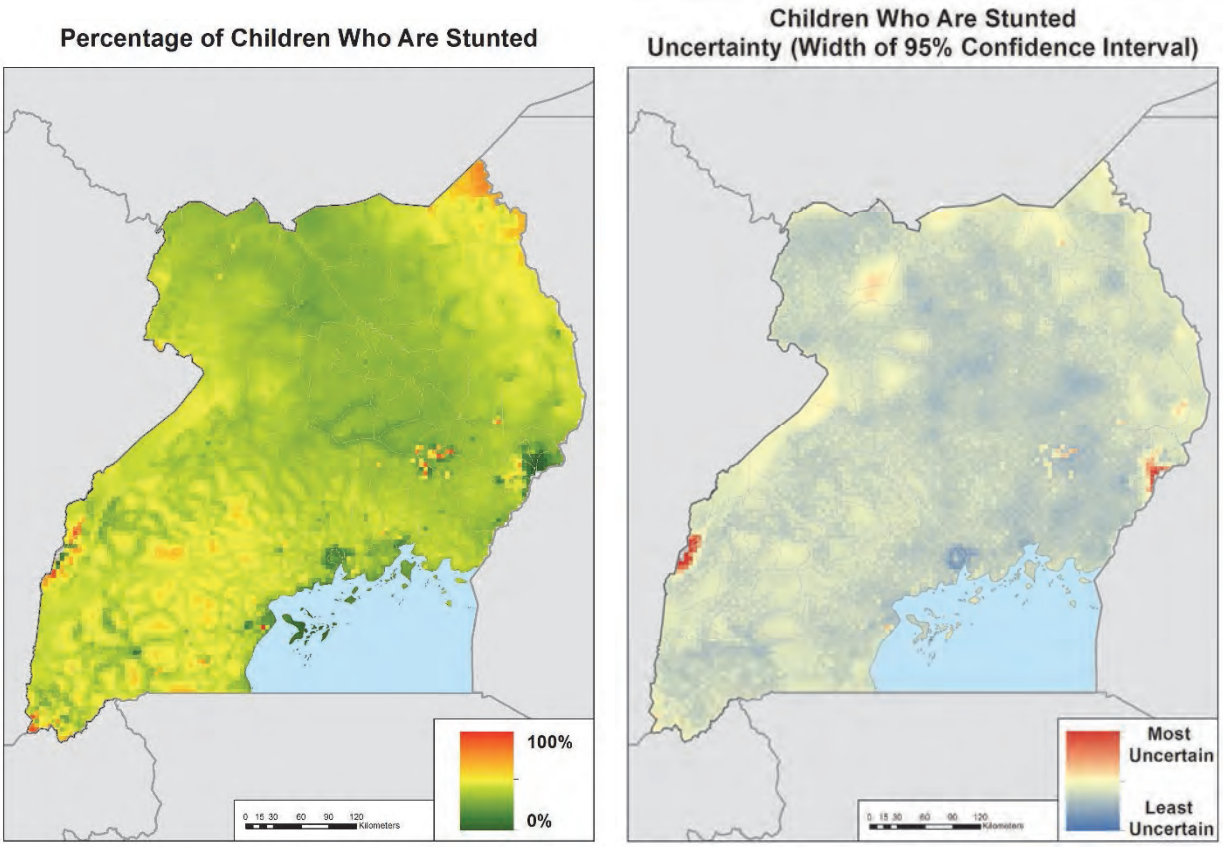


Figure 9. (left) Predicted map of indicator on proportion of children 6-59 months who are stunted in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.



3.2.3 Anemia prevalence in children

3.2.3.1 Exploratory analysis

Figure 10. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of children 6-59 months who are anemic.

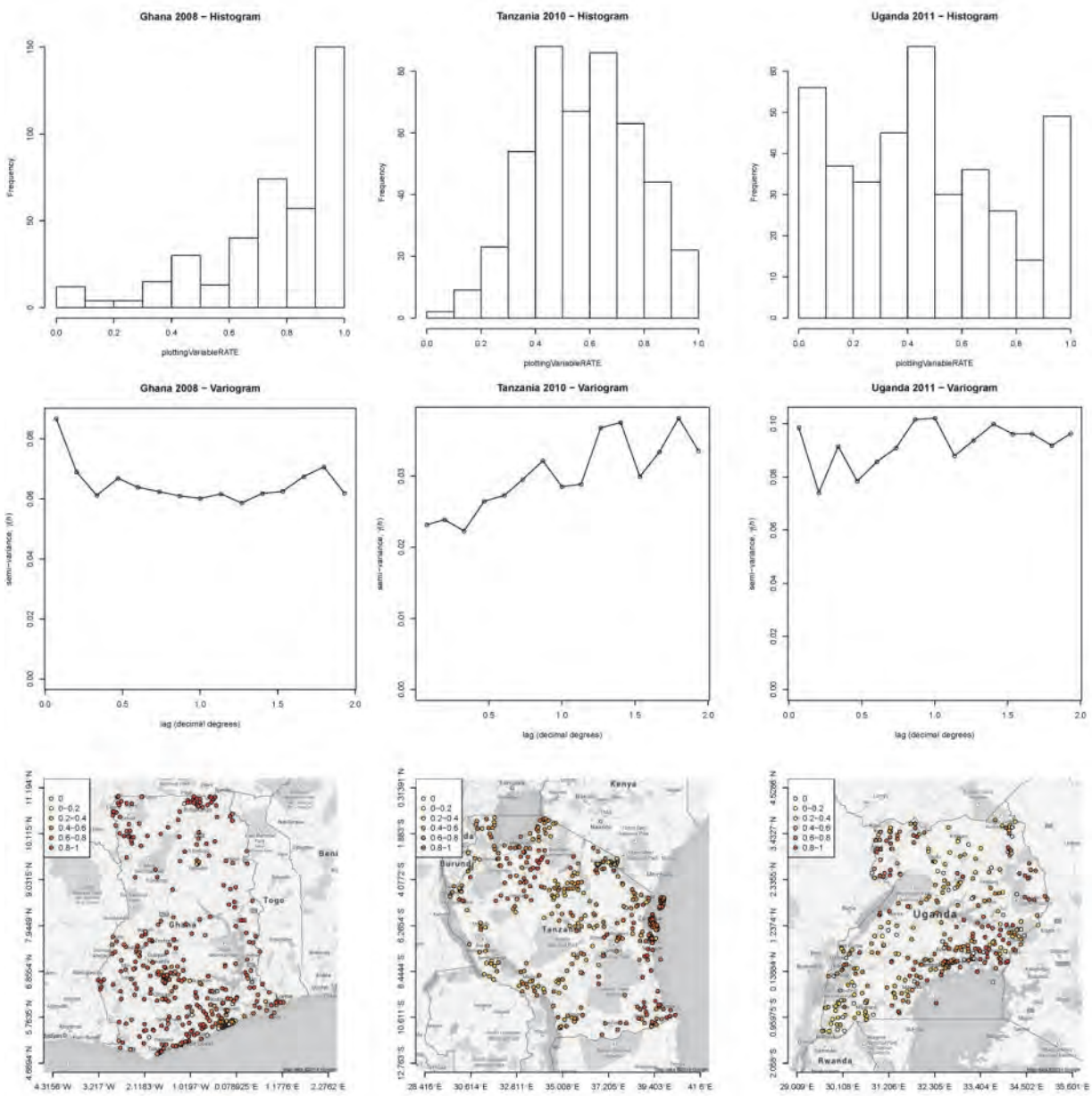


Figure 10 shows that the three study countries exhibit substantial differences in the proportions of surveys that show low and high levels of anemia, with rates apparently much higher in Ghana, very few areas of low prevalence in Tanzania, and a full range of proportions in Uganda. The variograms highlight that Tanzania shows the clearest levels of spatial dependence, although clustering of similar values are evident for Uganda and Ghana in the images.

3.2.3.2 Covariate selection

Table 6 summarizes the results of the covariate selection procedure for the indicator on proportion of children 6-59 months who are anemic. Shown for each country are details of the best-performing model, as defined by the exhaustive out-of-sample comparison of all individual and interaction terms across 17 initial candidate covariates. The covariate naming conventions and descriptions can be found in Table 1.

Table 6. Summary output from the covariate selection procedure for the indicator on proportion of children 6-59 months who are anemic. PR^2 is the predictive R-squared statistic. 'sqrt' denotes a square-root transform was applied to the covariate term.

	Ghana	Tanzania	Uganda
No. of Covariates	4	9	8
PR^2	0.100	0.259	0.150
Covariates in best model	sqrt(lights.2010) sqrt(aridity) **x** sqrt(lights.2010) sqrt(lights.2010) **x** sqrt(EVI) sqrt(lights.2010) **x** sqrt(TCB)	LST_night (elevation)^2 **x** sqrt(GPW) (elevation)^2 **x** (LST_day)^2 (elevation)^2 **x** LST (elevation)^2 **x** (TSI)^2 sqrt(PET) **x** LST_night (LST_day)^2 **x** (TSI)^2 TCB **x** (TSI)^2 TCW **x (**TSI)^2	sqrt(elevation) TCW sqrt(aridity) **x** sqrt(elevation) sqrt(elevation) **x** sqrt(PET) sqrt(elevation) **x** sqrt(precip) sqrt(GPW) **x** (EVI)^2 sqrt(grump.2010) **x** (EVI)^2 (LST_day)^2 **x** (TCB)^2 TCW **x** sqrt(precip

3.2.3.3 Geostatistical model: Validation statistics

Table 7 shows validation results from the geostatistical model for the indicator on proportions of children 6-59 months who are anemic. Again, low mean-square error values indicate the model is fitting with minimal overall bias. Mean absolute error values are lowest for Tanzania; this indicates this model performs with the greatest predictive precision.

Table 7. Summary validation results from the model-based geostatistical map for the indicator on proportion of children 6-59 months who are anemic. These are outputs from the four-fold out-of-sample validation exercise for each pilot country. MAE = mean absolute error, measuring the precision of predicted values; MSE = mean square error, measuring the bias of predicted values. Both metrics are in the same units as the variables themselves (which are all measured as proportions between zero and one).

	MAE	MSE
Ghana	0.157	0.0443
Tanzania	0.125	0.026
Uganda	0.210	0.070

3.2.3.4 Mapped surfaces

Table 6 shows elevation to be consistently selected as a covariate for the Tanzania and Uganda models where there are a wide range of cluster-level values to predict, as compared to Ghana, where the vast majority of clusters showed high levels of anemia prevalence. Despite accounting for elevation in the calculation of anemia rates, elevation remained as a key covariate for Tanzania and Uganda, and is reflected in the output prediction maps in

Figure 12-13, where the highest rates are predicted to be in the areas of high elevation. Like the stunting indicator, Ghana was predicted with least uncertainty. Across all four indicators, the levels of uncertainty for anemia prevalence in Tanzania and Uganda were the highest, and are likely driven by the high levels of heterogeneity in anemia prevalence in those countries.

Figure 11. (left) Predicted map of indicator on proportion of children 6-59 months who are anemic in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.

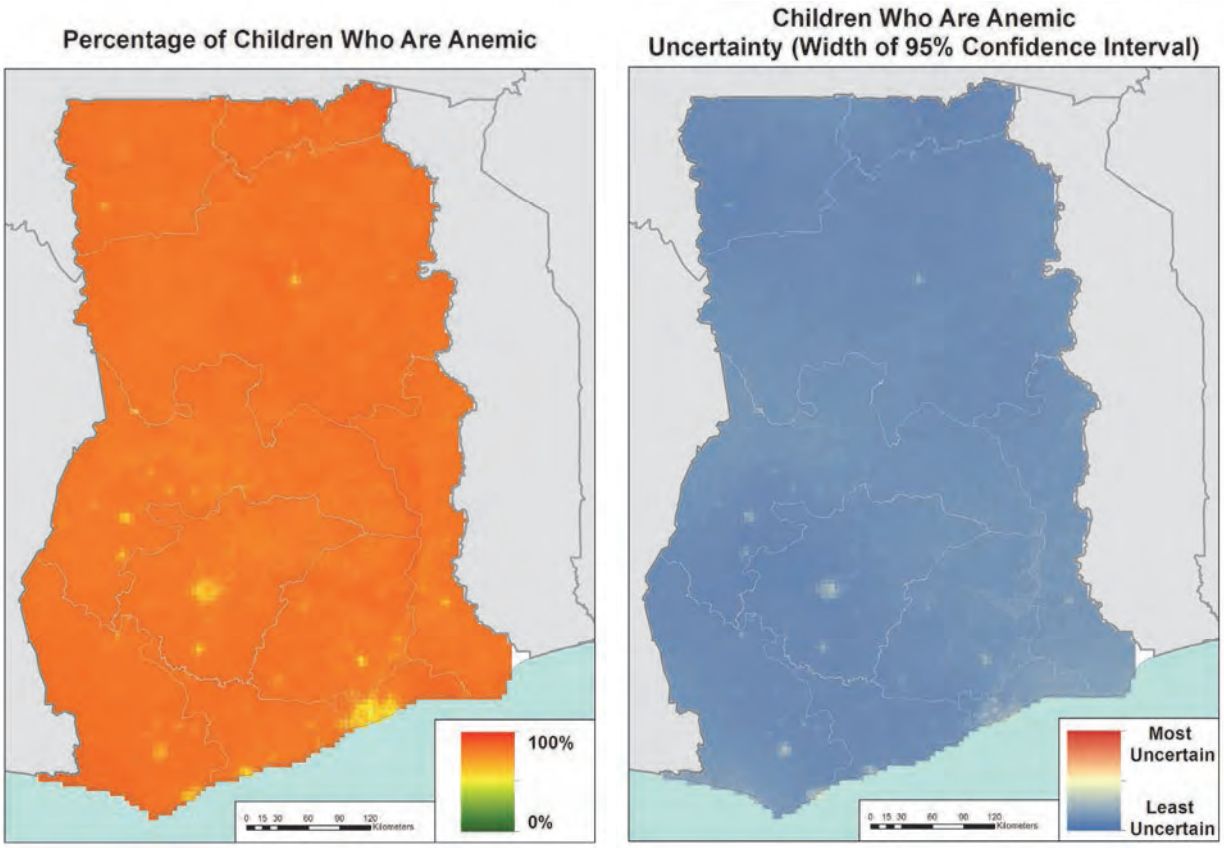


Figure 12. (left) Predicted map of indicator on proportion of children 6-59 months who are anemic in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.

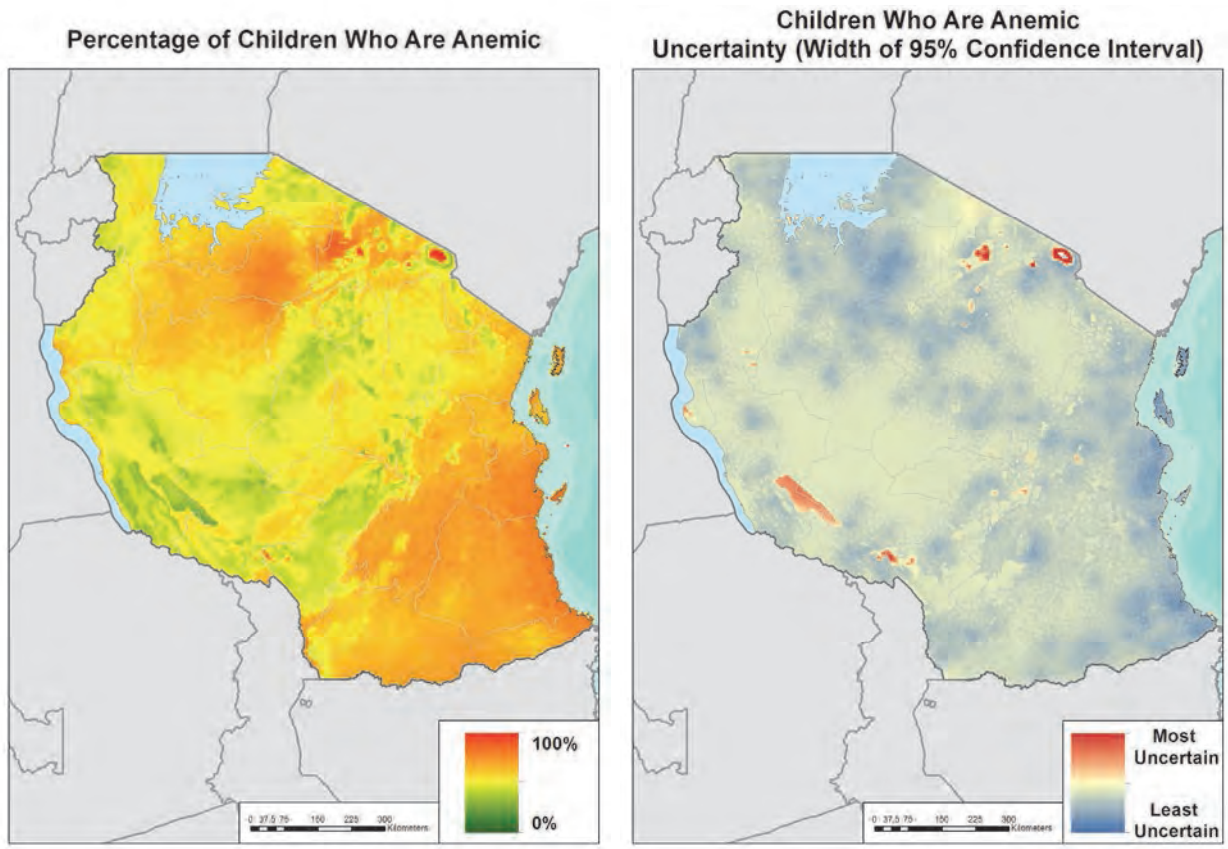
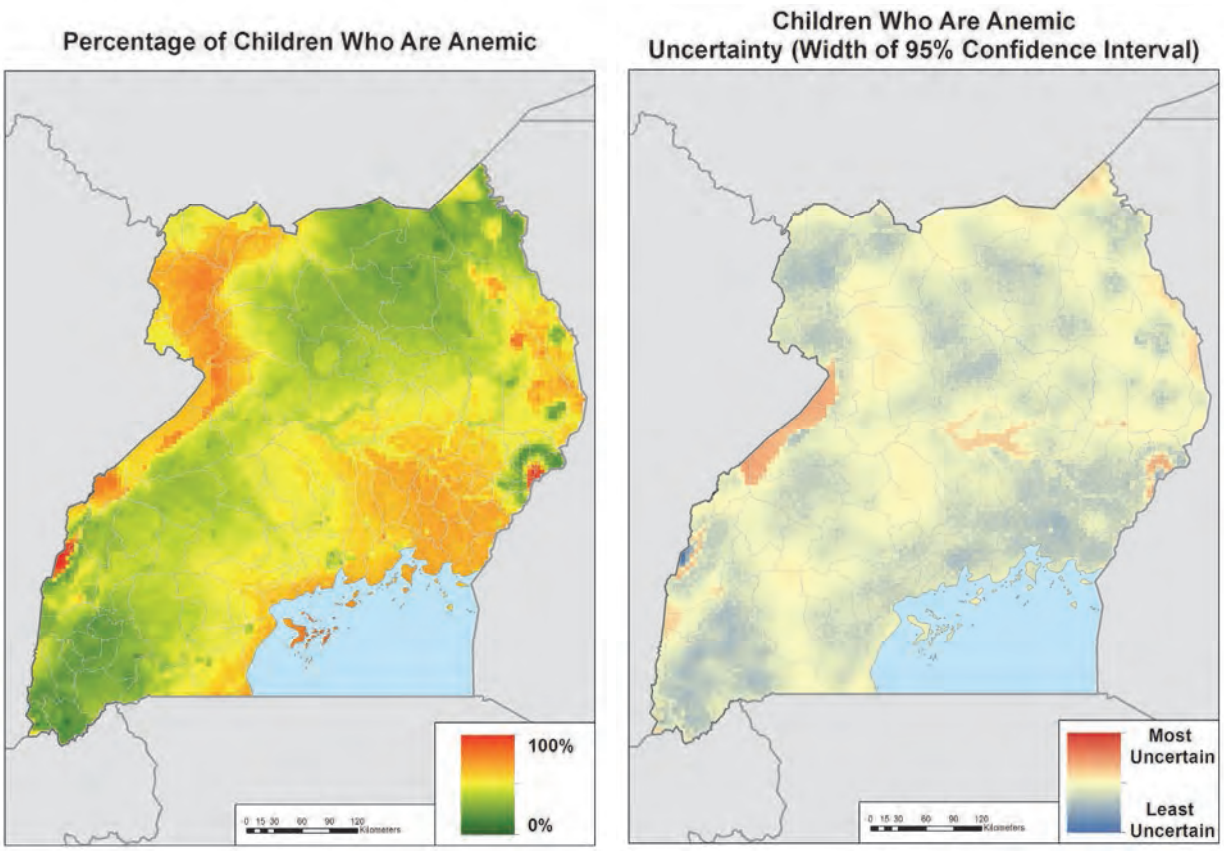


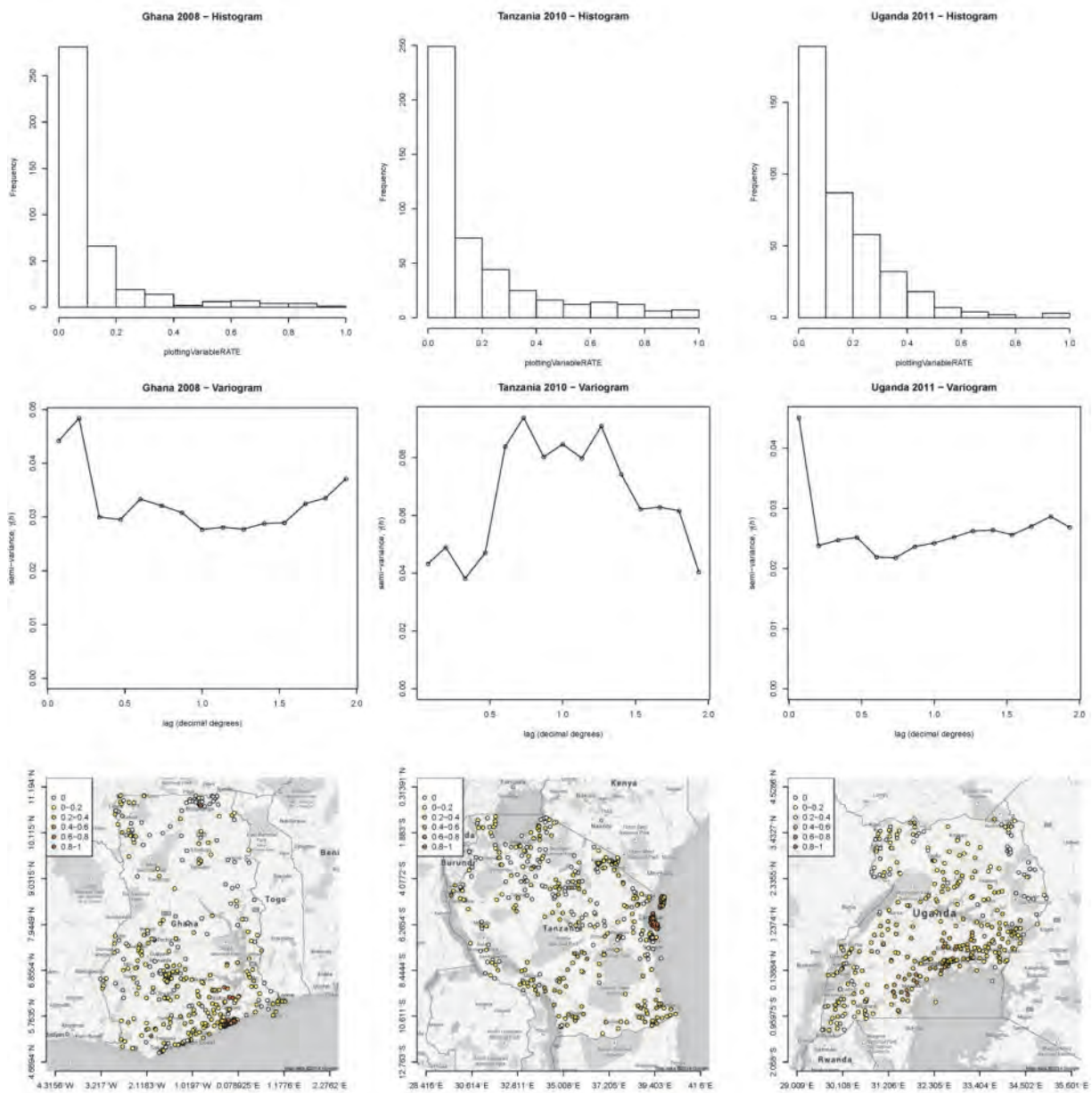
Figure 13. (left) Predicted map of indicator on proportion of children 6-59 months who are anemic in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.



3.2.4 Access to improved sanitation

3.2.4.1 Exploratory analysis

Figure 14. Histograms (top row), variograms (middle row), and cluster-level survey data (bottom row) for the indicator on proportion of households with improved sanitation.



All three countries show similar distributions of survey outcomes, with the vast majority of clusters having zero or low access to improved sanitation. For each of the countries, almost all clusters with high proportions of households having access to improved sanitation are in urban areas (with the exception of rural areas of Zanzibar), although substantial heterogeneity within the major urban areas is evident.

3.2.4.2 Covariate selection

Table 8 summarizes the results of the covariate selection procedure for the indicator on proportion of households with improved sanitation. Shown for each country are details of the best-performing model, as defined by the exhaustive out-of-sample comparison of all individual and interaction terms across 17 initial candidate covariates. The covariate naming conventions and descriptions are found in Table 1.

Table 8. Summary output from the covariate selection procedure for the indicator on proportion of households with improved sanitation. PR^2 is the predictive R-squared statistic. 'sqrt' denotes a square-root transform was applied to the covariate term.

	Ghana	Tanzania	Uganda
No. of Covariates	5	12	8
PR^2	0.168	0.547	0.234
Covariates in best model	lights.2010 sqrt(lights.2012) sqrt(afripop.2008) **x** sqrt(lights.2012) lights.2010 **x** sqrt(PET) sqrt(lights.2012) **x** sqrt(PET)	(TSI)^2 sqrt(access) **x** LST aridity **x** (LST_night)^2 aridity **x** (TSI)^2 sqrt(GPW) **x** (precip)^2 sqrt(grump.2010) **x** (TSI)^2 sqrt(lights.2010) **x** (precip)^2 sqrt(PET) **x** LST sqrt(PET) **x** (TSI)^2 (EVI)^2 **x** (TSI)^2 (LST_day)^2 **x** (TSI)^2 (LST_night)^2 **x** (TSI)^2	(PET)^2 sqrt(access) **x** (elevation)^2 sqrt(access) **x** sqrt(lights.2010) sqrt(access) **x** (PET)^2 sqrt(access) **x** (LST_night)^2 sqrt(access) **x** sqrt(precip) sqrt(GPW) **x** (EVI)^2 sqrt(grump.2010) **x** (EVI)^2

3.2.4.3 Geostatistical model: Validation statistics

Table 9 shows validation statistics from the geostatistical model for the indicator on the proportion of households with improved sanitation. As well as confirming an unbiased model fit (small MSE), the precision measures (MAE) are also generally smaller for this indicator than the other three investigated, particularly for Ghana and Uganda.

Table 9. Summary validation results from the model-based geostatistical map for the indicator on proportion of households with improved sanitation. These are outputs from the four-fold out-of-sample validation exercise for each pilot country. MAE = mean absolute error, measuring the precision of predicted values; MSE = mean square error, measuring the bias of predicted values. Both metrics are in the same units as the variables themselves (which are all measured as proportions between zero and one).

	MAE	MSE
Ghana	0.083	0.018
Tanzania	0.111	0.025
Uganda	0.086	0.013

3.2.4.4 Mapped surfaces

Access to improved sanitation produced the overall highest predictive R^2 values, with up to 0.55 for Tanzania, and relatively precise, unbiased models. The mapped surfaces in Figure 15-17 show the significant rural-urban divides that exist across the three countries in terms of access to sanitation, driven by the access and population density covariates included through the covariate selection process. Across all three countries, uncertainty was almost uniformly low, with the exception of isolated pockets of higher uncertainty, particularly in Tanzania.

Figure 15. (left) Predicted map of the indicator on proportion of households with improved sanitation in Ghana. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.

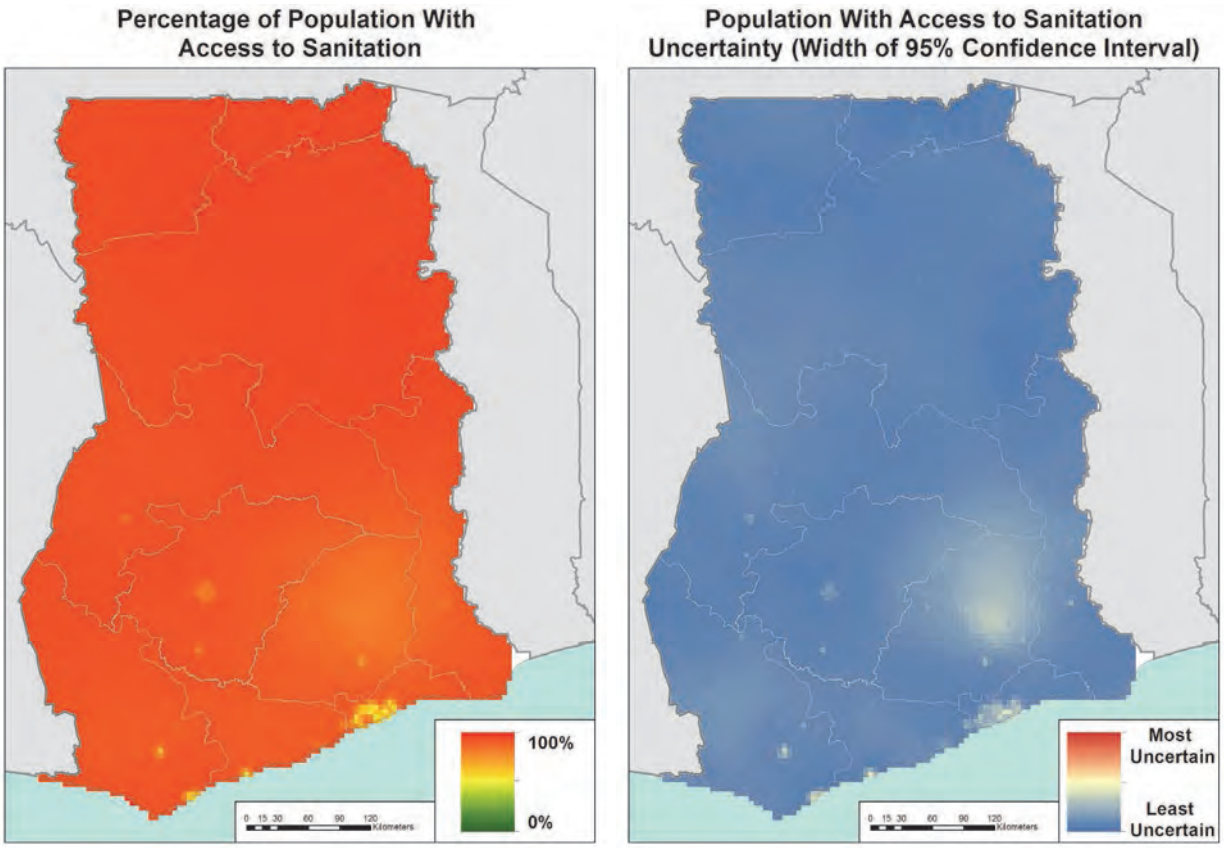


Figure 16. (left) Predicted map of the indicator on proportion of households with improved sanitation in Tanzania. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.

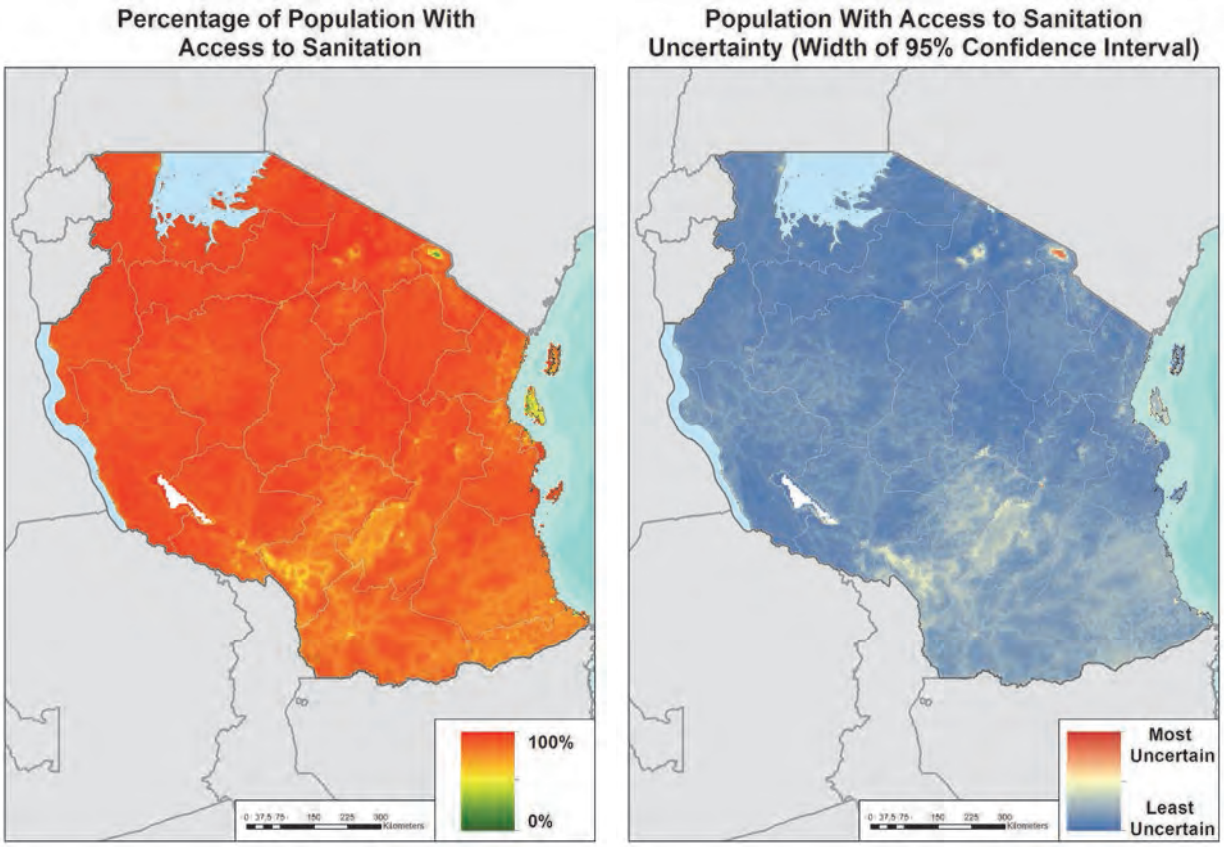
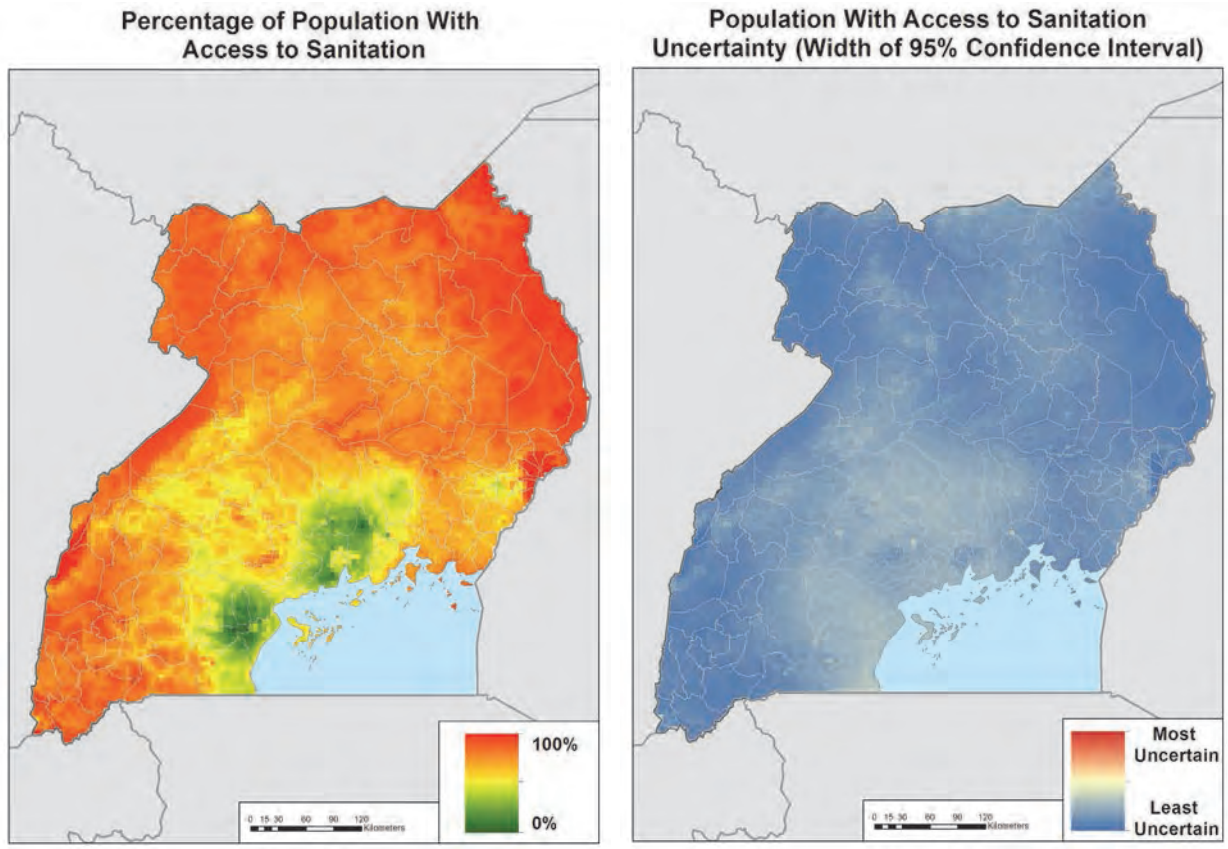


Figure 17. (left) Predicted map of the indicator on proportion of households with improved sanitation in Uganda. The continuous surface is the posterior mean prediction at 5x5km resolution. (right) Accompanying map of model-based uncertainty. The uncertainty metric is the width of the 95% credible interval with low uncertainty in blue and high uncertainty in red.



4. Effects of Cluster Centroid Displacement

4.1 Methods

4.1.1 Overview

The impact of centroid displacement was investigated in three stages. First, a realistic set of displaced data was generated for use in testing. Second, the impact of that displacement on descriptive statistical properties of the DHS data was explored. Third, the effect of displacement on the precision of a full Bayesian model-based geostatistical interpolation was explored. These three methodological stages are described in detail in the following sections.

4.1.2 Generating displaced data

As discussed, all DHS survey data are subject to a standardized displacement procedure to protect respondents' anonymity, and to respect the more sensitive aspects of the questionnaires such as HIV infection status. The displacement procedure is as follows:

- Urban clusters are displaced a distance up to two kilometers.
- Rural clusters are displaced a distance up to five kilometers, with a further, randomly selected 1% of the rural clusters displaced a distance up to ten kilometers.

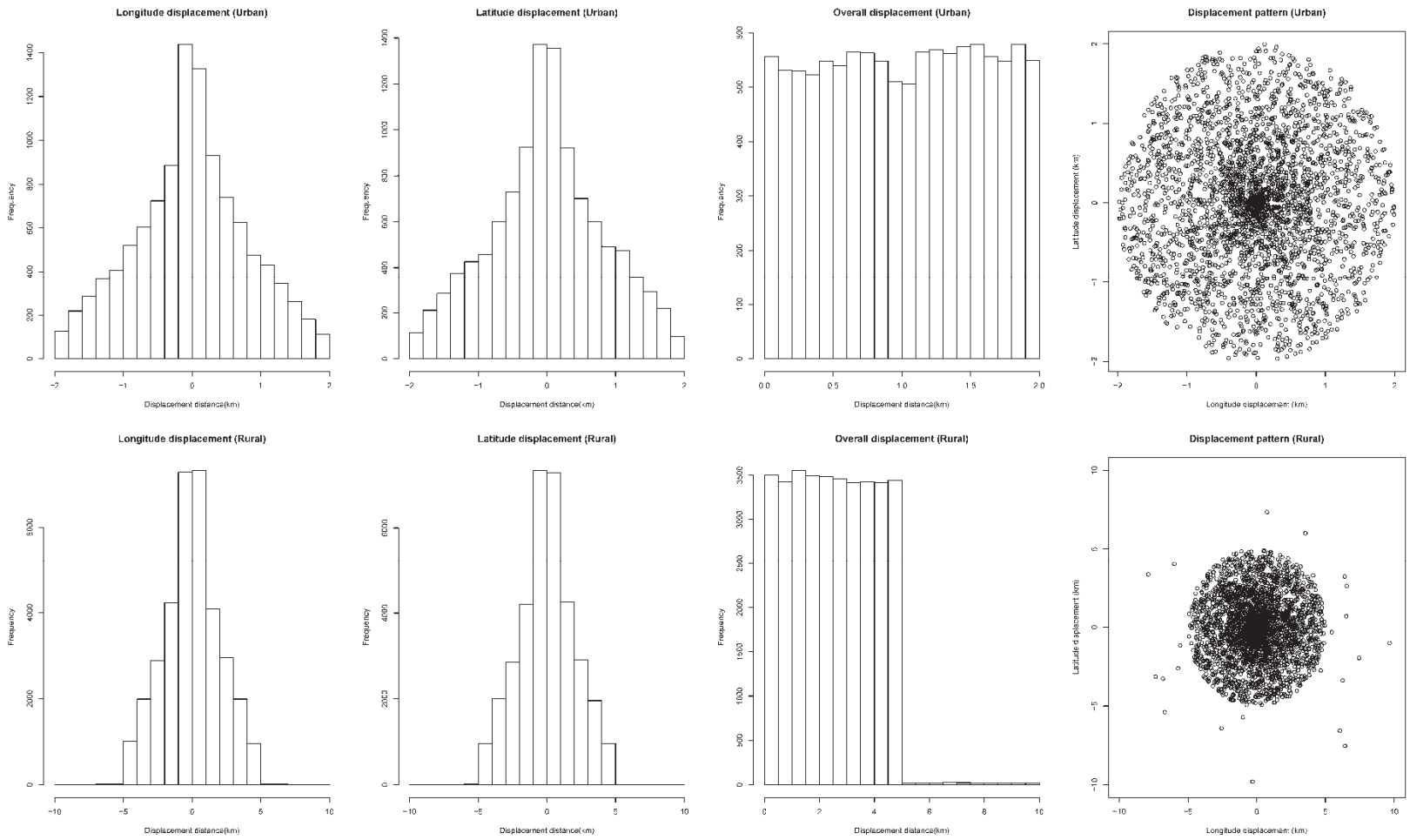
The displacement of GPS data is restricted so that the coordinates remain within the original country and DHS survey region. For surveys after 2008, the displacement is further restricted to the second administrative level (named "Districts" in most countries) when reliable boundary data are available. Details on the DHS georeferenced data displacement process and the spatial variability of the resulting data can be found in DHS Spatial Analysis Report 7 (Burgert et al. 2013).

In this study, we recreated the DHS displacement algorithm and applied it to non-displaced versions of the survey cluster coordinates for the three surveys: the Ghana 2008 DHS, Tanzania 2010 DHS, and Uganda 2011 DHS. Within the constraints outlined above, the displacement is stochastic and uses a random number generator to specify the angle and distance of the displacement. The algorithm was run 100 times to provide a sufficiently large set of stochastic realizations of the displacements; this allowed subsequent exploration of the likely range of impact these have on downstream analyses.

Following the creation of the 100 displaced sets for each survey, a set of diagnostic checks were used to ensure the displacements displayed the expected characteristics.

Figure 18 illustrates these characteristics for one exemplar survey (the Tanzania 2010 DHS) and shows the random displacement distances for urban and rural points, and the two-dimensional pattern of displacements relative to the true centroid.

Figure 18. Example set of randomly displaced points implemented according to the standardized DHS displacement algorithm using the 2010 Tanzania DHS. The histograms (left column) show the distribution of displacement distances and the point maps (right column) plot the location of each randomly displaced point relative to the true centroid in the center of the image. The top row relates to urban points and the bottom row to rural points. Note that the urban and rural displacement pattern maps are plotted on different distance scales.



4.1.3 Exploration of effect of displacement on statistical properties of data

After generating the 100 randomly displaced sets for each survey, we explored how displacement affected the basic characteristics of the survey variables.

4.1.3.1 Impact on spatial autocorrelation structure

First, the impact of displacement on spatial autocorrelation structure was investigated. Achieving acceptable precision with any geostatistical interpolation exercise is contingent on data displaying spatial structure, i.e., the tendency for nearby locations to display more similar properties than those further apart. Depending on the spatial scales of variation present in each indicator, it was possible that randomly displacing the observation locations would act to degrade or destroy that short-scale spatial autocorrelation, with important consequences for the feasibility of subsequent geostatistical prediction. To test this, empirical spatial variograms were computed for each of the 100 displaced sets, each of the three countries, and each of the four indicators described earlier. For each country and variable, the 100 displaced-point variograms were then combined and compared to the single non-displaced variogram.

4.1.3.2 Impact on relationship with environmental covariates

A second important requirement for optimal geostatistical mapping is that the observed data are correlated with underlying covariates that can then be used in a multivariate fixed-effect component of a geostatistical model. To test this, the following procedure was implemented for each country and indicator:

1. For each of the environmental covariates described in Section 5.1.3.1, values were extracted at the locations of the non-displaced cluster data.
2. For each indicator, a multivariate generalized linear model was then fitted to the full set of covariate values to predict the response indicator.
3. The out-of-sample (predictive) R^2 value was calculated as a simple summary of the regression model predictive performance.
4. Steps 1-3 were then repeated for each of the 100 randomly displaced sets.

Histograms were then generated that show the distribution of R^2 values across the 100 displaced-point regressions relative to the non-displaced reference model.

4.1.4 Exploration of effect of displacement on MBG-derived interpolated surfaces

Section 5.1.5 describes the structure and implementation of Bayesian model-based geostatistical mapping of the four indicators, based on the publically available, displaced survey data from the Uganda 2011, Tanzania 2010, and Ghana 2008 DHS surveys. To test the nature and magnitude of impacts on predictive accuracy induced by the displacement procedure, the entire geostatistical procedure was re-run in full for each of the 100 randomly displaced sets for each indicator and across all three surveys; there was an additional run on the non-displaced reference data in each case. This means a total of 1,200 models were built and implemented. In each case, an out-of-sample validation procedure was run in parallel across the 100 sets and for the single non-displaced set.

In the main mapping exercise described earlier (see section 5.1.5.3), the validation statistics computed for comparison were the correlation coefficient (between actual and predicted values), the predictive R^2, the MSE and the MAE. The distribution of each of these summary validation metrics across the 100 sets was

plotted as a histogram for each indicator, along with a reference line that denoted the performance of the corresponding non-displaced set.

To explore the impact of the displacements on the mapped surfaces, each of the 100 maps were compared and, for each predicted pixel, the variation in the predictions induced by the displacements was summarized by their standard deviation. This value was then mapped to allow geographic comparison in the effect of displacement.

4.2 Results

4.2.1 Generating displaced data

Figures 19-21 show the locations of the 100 sets of displaced centroids relative to the non-displaced sets for the Ghana 2008, Tanzania 2010, and Uganda 2011 DHS surveys, respectively. These point maps provide a useful visualization of the magnitude of the displacement effect relative to the geographical scale of the country and the separation distances between centroids.

Figure 19. Location of the 100 randomly displaced cluster centroid from the 2008 Ghana DHS (black dots). Shown for reference in the center of each is the original non-displaced location, shown as either a red dot (urban centroids) or green dot (rural centroids).

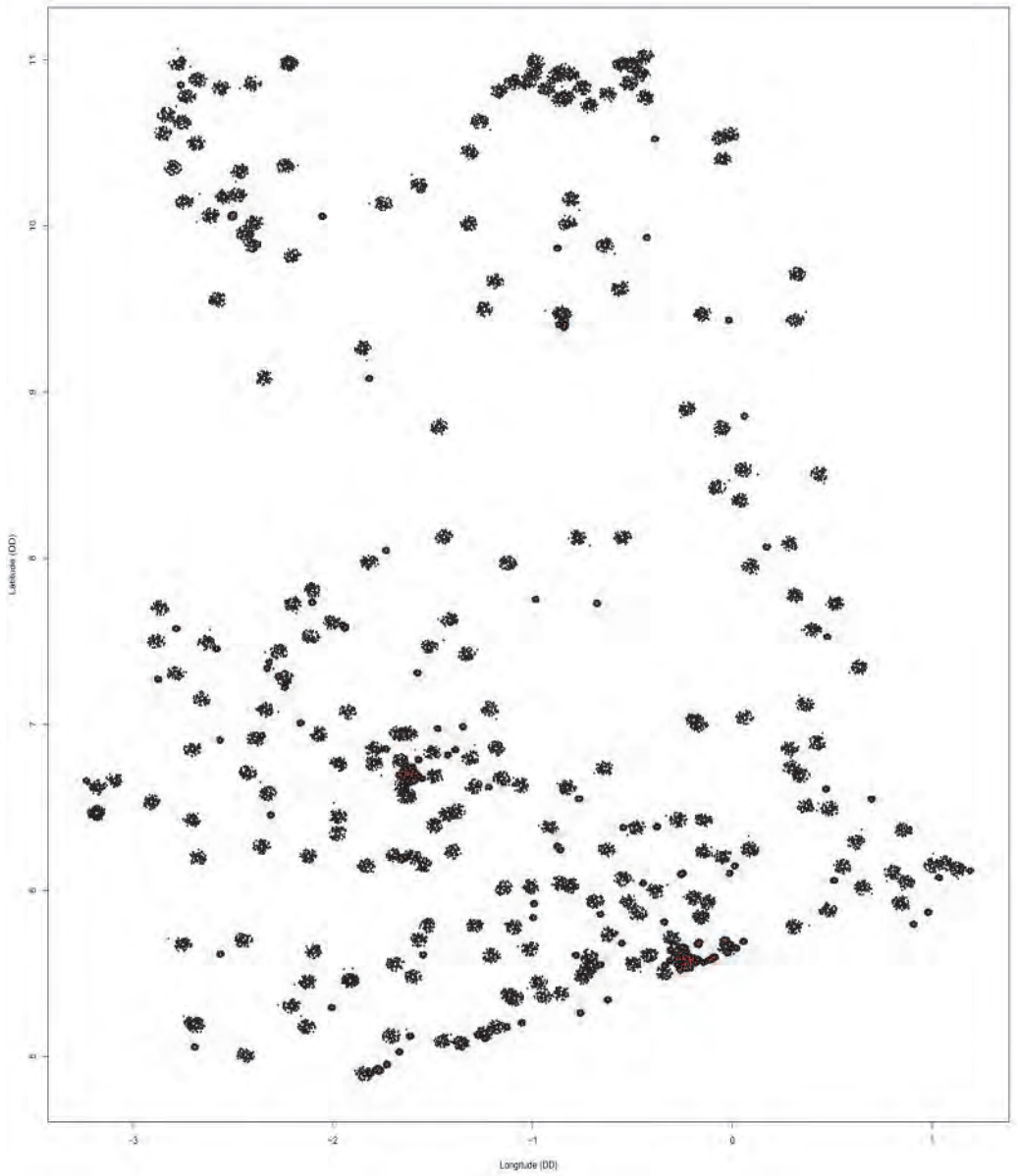


Figure 20. Location of the 100 randomly displaced cluster centroid from the 2010 Tanzania DHS (black dots). Shown for reference in the center of each is the original non-displaced location, shown as either a red dot (urban centroids) or green dot (rural centroids).

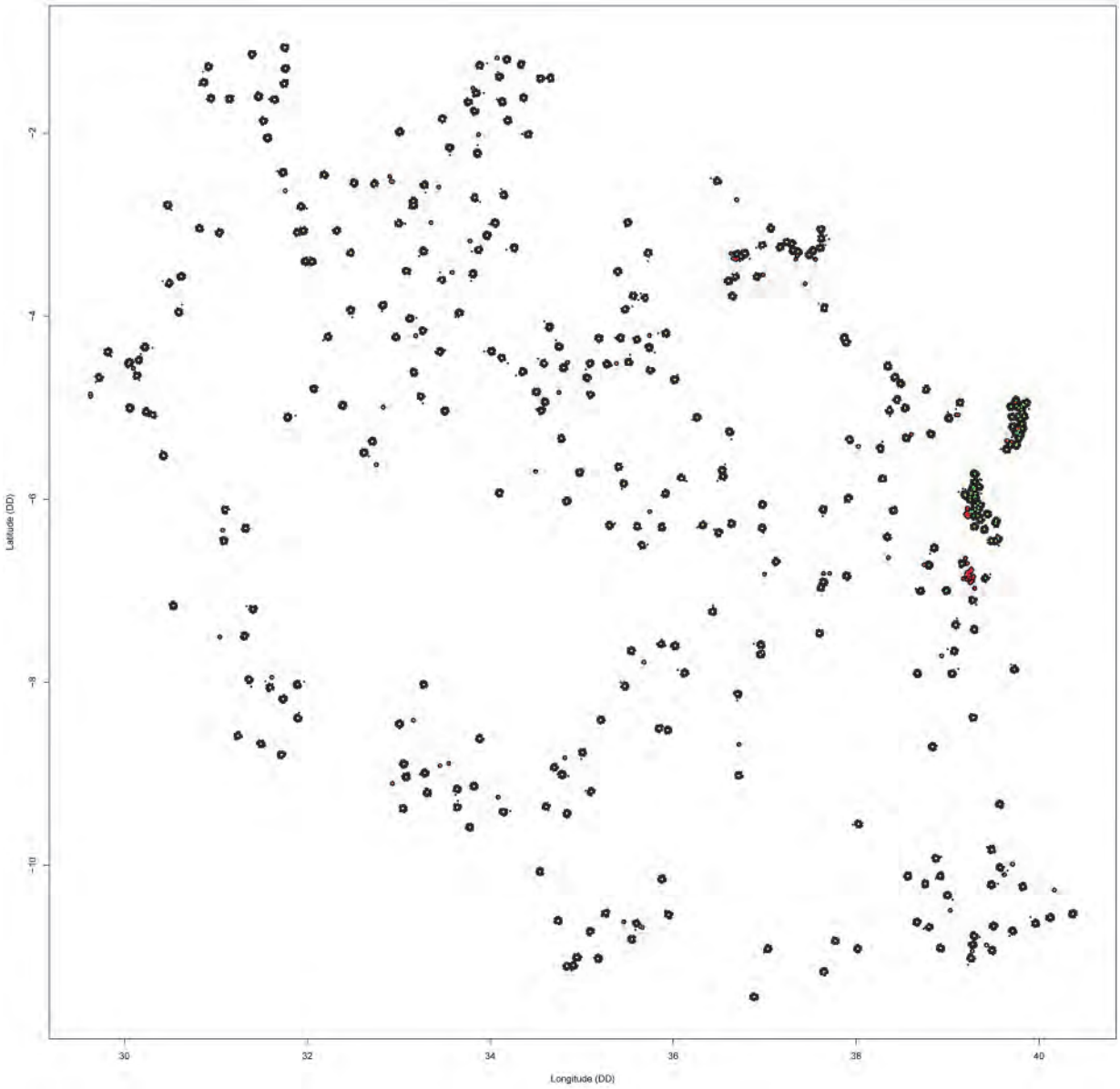
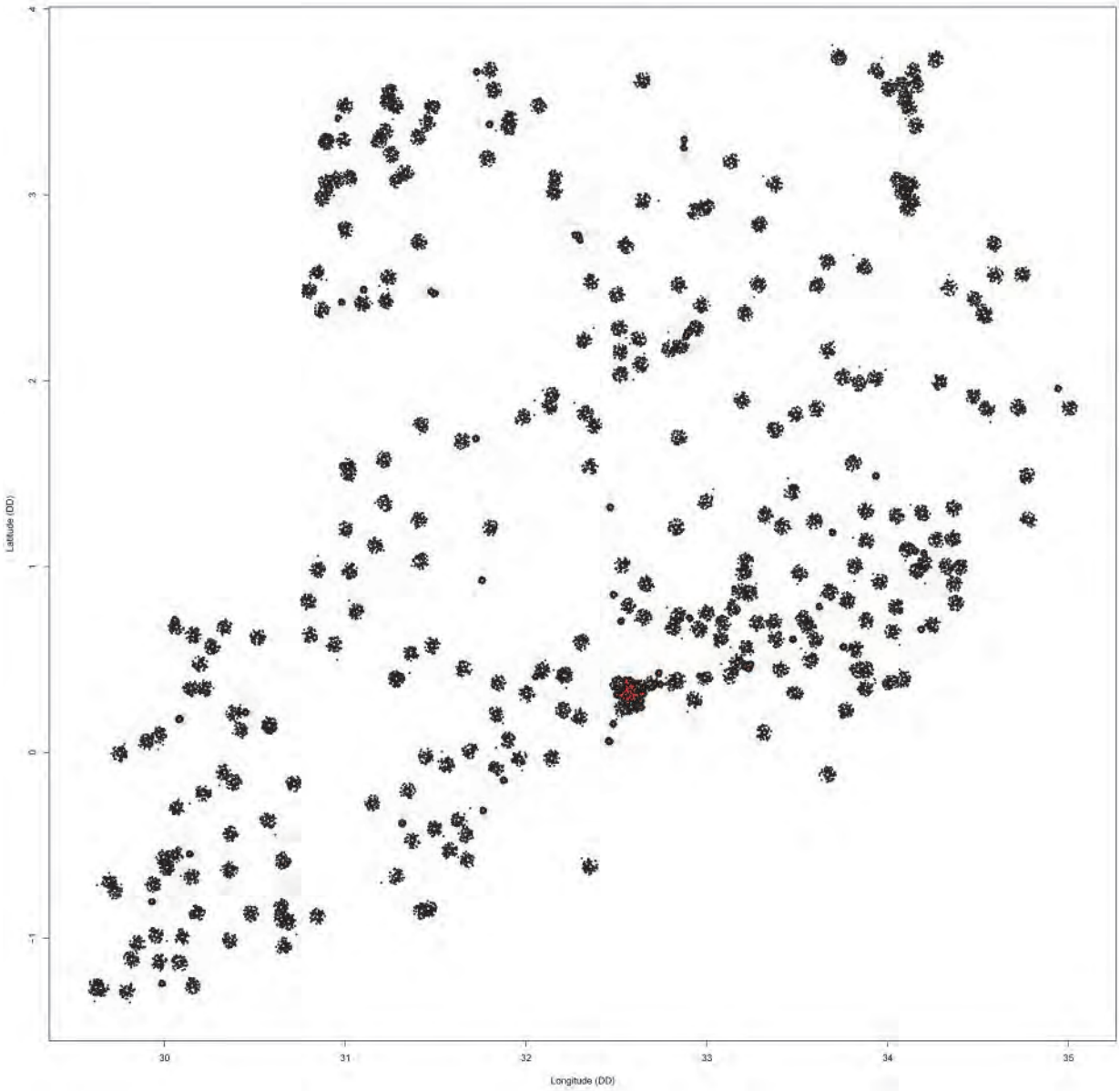


Figure 21. Location of the 100 randomly displaced cluster centroid from the 2011 Uganda DHS (black dots). Shown for reference in the center of each is the original non-displaced location, shown as either a red dot (urban centroids) or green dot (rural centroids).

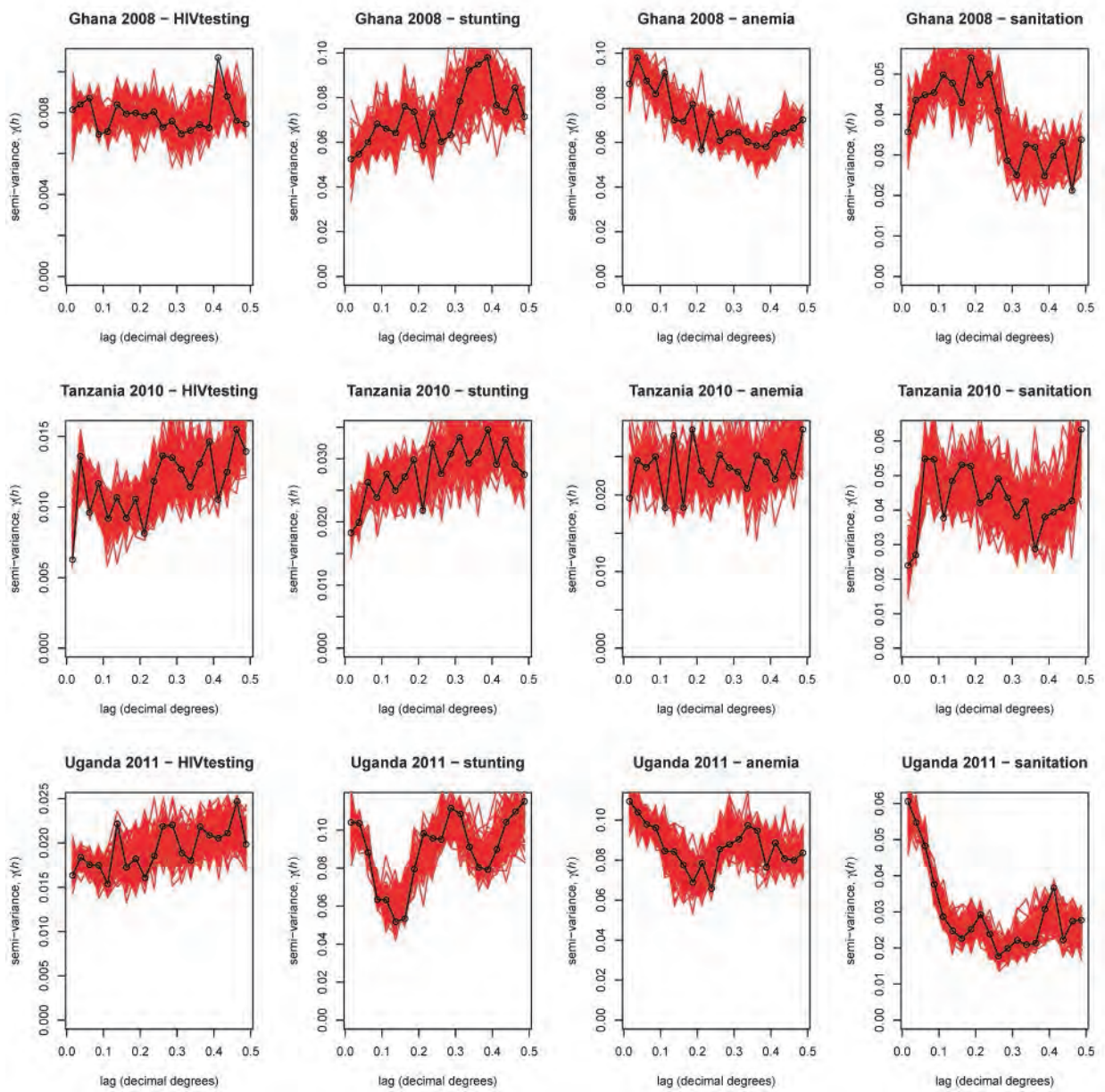


4.2.2 Exploration of effect of displacement on statistical properties of data

4.2.2.1 Impact on spatial autocorrelation structure

Figure 22 shows the results from the variogram analysis, which depict for each country-indicator the variation in empirical variograms across the 100 displaced sets relative to the non-displaced reference set. In general, the displaced variograms do not differ markedly from the non-displaced reference set. This suggests that the autocorrelation structure is relatively robust to degradation induced by the displacement process. However, one important caveat is that, in many cases, the non-displaced data themselves have only relatively weak structure, as indicated by the small structured-to-unstructured variance ratio. Reassuringly, however, for those country-indicators with more pronounced structure (e.g., the sanitation indicator from the Tanzania 2010 survey), the displaced variograms retain the primary features with similar magnitude nugget, range, and sill features. A second important caveat is that, intuitively, the effect of displacement will become progressively more important as shorter scales of variation are considered, and structure will inevitably be broken down when considering variation within the displacement radius itself. However, for the purposes of generating national-scale interpolated surfaces, these micro-scales of variation are less important than the larger structures captured in these variograms, which appear to be relatively robust and suggest that displacement will not dramatically limit the precision of geostatistical models. This conclusion may not be valid when discussing mapping within urban areas, since in those geographically constrained and highly heterogeneous environments, capturing the micro-scale variation is more important. The specific challenges of mapping with displaced data in urban areas is discussed elsewhere in this report.

Figure 22. The effects of centroid displacement on empirical variograms structure. Each panel shows the set of 100 variograms (red lines) derived from each displaced cluster set, overlaid with the variogram for the non-displaced set (black line). Panels are organized by survey (rows) and indicators (columns). 0.5 decimal degrees is approximately 55 km.



4.2.2.2 Impact on relationship with environmental covariates

For each country-indicator, Figure 23 shows a histogram of predictive R^2 values relating to regression models described in Section 6.1.3.2, derived from the 100 displaced centroid sets, with the corresponding R^2 value for the non-displaced set shown for reference. Predictive R^2 summarizes the power of the regression model to explain variation in the indicator, when tested out-of-sample (formally, the proportion of variation in the response data explained by the multivariate regression model - so larger R^2 values reflect superior model performance.) Comparison of each histogram with the reference (red) line in each case

reveals varied results. Of the twelve country-indicators, six show the majority of displaced regression models performing worse than the non-displaced, i.e., the bulk of the histogram density is to the left of the red line (Tanzania HIV testing, stunting, anemia, sanitation; Ghana anemia; Uganda sanitation); five show no clear effect, i.e., the red line is approximately central on the histogram (Ghana HIV testing, stunting; Uganda HIV testing, stunting, anemia); while one apparently shows the displaced model performing, on average, better than the non-displaced, i.e., the histogram density mainly to the right of the red line (Ghana sanitation). The *a-priori* expectation is that displacement should have a detrimental effect of some magnitude, because it artificially dislocated any relationships that are present between the response indicator and the environmental covariate. The most plausible explanation for the mixed results seen here is that, for many of these models, the multivariate relationship between the covariates and the indictor is already relatively weak for the non-displaced model. Thus, displacing the points and dislocating the linkage between independent and dependent variables has little measurable effect or, in some case, results in a small improvement which is essentially a stochastic effect. Importantly, when only those non-displaced models with relatively better predictive performance are considered (e.g., Tanzania-sanitation; Uganda-sanitation; Tanzania-stunting; Tanzania-HIV testing), all display the expected result of substantial degradation on performance in the displaced sets relative to the non-displaced.

4.2.3 Exploration of effect of displacement on MBG-derived interpolated surfaces

4.2.3.1 Effect on predictive validation statistics

The plots shown in Figures 24-27 summarize the effects of displacement on the predictive accuracy of model-based geostatistical interpolation of the four indicators. Each histogram shows the distribution of one of the four validation statistics for models run across each of the 100 displaced sets, with the statistic value from the non-displaced set shown for reference. In a similar way to the non-spatial regression model results described in section 6.2.2.2, these results are complex and describe a range of different outcomes depending on the country and indicator in question although, behaviors *between* the four alternative validation statistics tends to be broadly consistent. For example, where out-of-sample correlation between actual and predicted values appears to be substantially degraded by displacement, the predictive R^2 values also tend to be systematically smaller across the displacement sets, and the measures of model bias and precision (MSE and MAE, respectively) tend to be larger. Similarly, where correlation or R^2 show no strong effect of displacement (where the reference non-displaced red line lies centrally in the histogram density), there tends to be little effect for MSE or MAE.

As in the non-spatial regression analysis, the most pronounced effects of displacement are seen in those country-indicators that are relatively well predicted for the non-displaced set; where that original prediction is poor, the effect of displacement are less pronounced. There is no strong pattern across the different countries and indicators as to which are consistently well predicted and thus more noticeably affected by displacement. Figure 24 shows the results for the indicator on HIV testing rates. For this indicator, displacement has a weak but noticeable effect in the case of the Ghana 2008 DHS data: correlation and R^2 both tend to be smaller for displaced sets relative to non-displaced, and MAE and MSE are correspondingly higher. For the Tanzania 2010 DHS data, displacement appears to have no consistent effect across any of the statistics, with the reference line central across all. For the Uganda 2011 data, displacement appears to slightly improve the statistics, with higher correlations and R^2 and lower MSE and MAE.

Figure 23. The effects of centroid displacement on indicator relationships with gridded covariates. Each panel shows a histogram of predictive R^2 from 100 multivariate linear regression models based on the 100 displaced sets. The red vertical line demonstrates the R^2 value for the model based on the non-displaced set. Panels are organized by indicator (rows) and survey (columns).

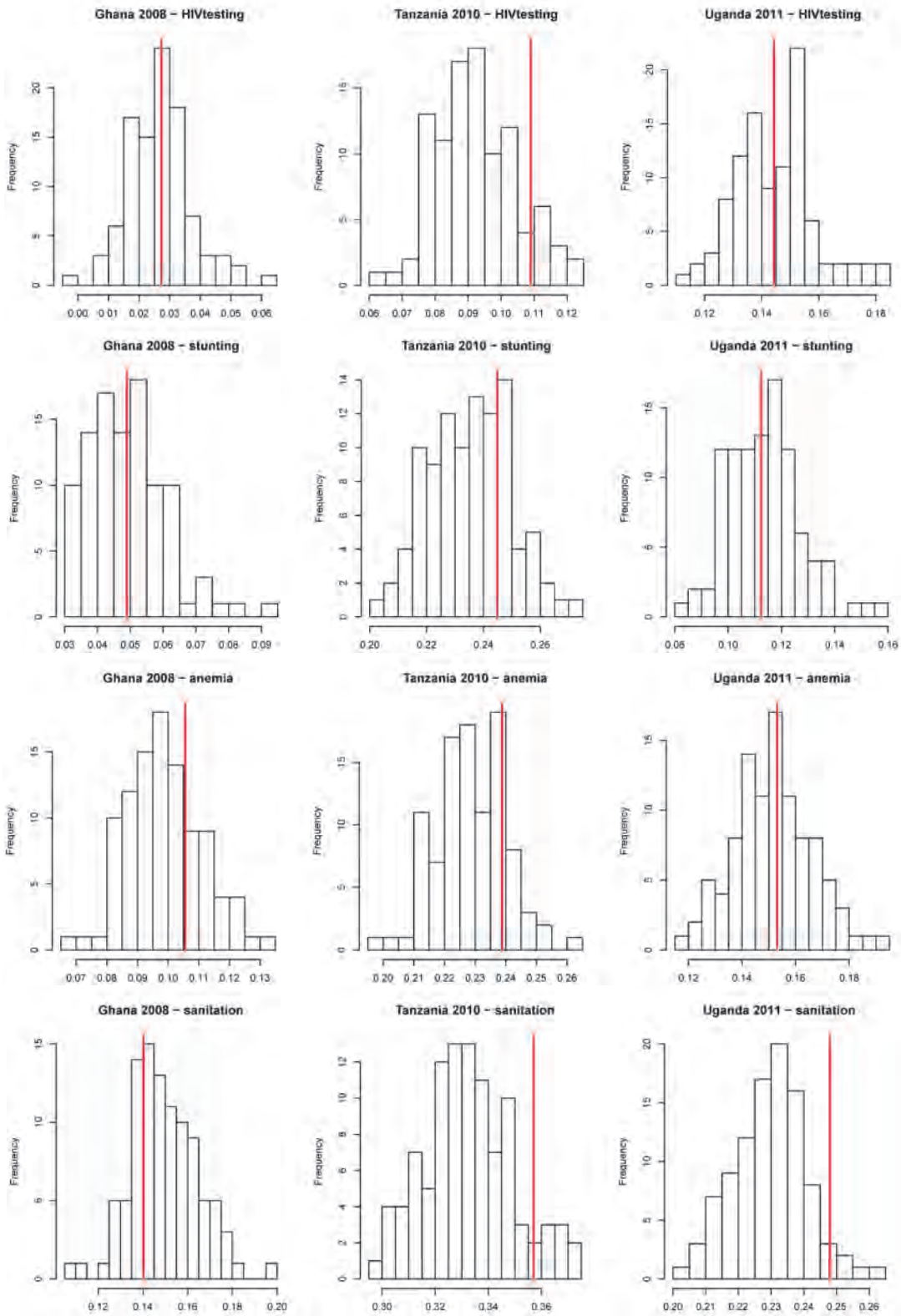


Figure 25 shows results for the indicator on prevalence of stunting. Here, the non-displaced model performs relatively well for the Tanzania 2010 and Uganda 2011 data, and displacement has a small but noticeable detrimental effect across the validation statistics. However, with Ghana 2008 data, the non-displaced model performs poorly and, as expected, displacement has little additional effect. Figure 26 shows results for the indicator on anemia prevalence in children. Interestingly, the impact of displacement appears most pronounced for the Ghana 2008 data, despite the relatively weak performance of the non-displaced model. The impact is far less pronounced for the Tanzania and Uganda data, despite the non-displaced models being more predictive in these countries. Figure 27 shows results for the indicator on access to improved sanitation. The Ghana 2008 and Tanzania 2010 results demonstrate significant displacement impact across all four statistics, particularly in Tanzania. Impacts are more moderate for the Uganda 2011 data.

Given the relatively mixed results, drawing general conclusions is challenging. Where the non-displaced model performs weakly, displacement neither improves nor dramatically hinders the performance of the model; the outcome remains an interpolated map with relatively imprecise pixel-level predictions. Where the original non-displaced model performs more strongly, the effects of displacement consistently reduce both the precision and unbiasedness of the resulting map. Importantly, the magnitude of this detrimental effect tends to be relatively small.

Figure 24. Effect of displacement on performance of model-based geostatistical predictive performance for the HIV testing indicator. Each histogram shows the distribution of one of four validation statistics (Cor = correlation between observed and predicted; R2 = predictive R^2; MSE = mean square error; MAE = mean absolute error) for models based on the 100 displaced sets relative to the non-displaced reference set (red vertical line). Panels are organized by validation statistic (rows) and DHS survey (columns).

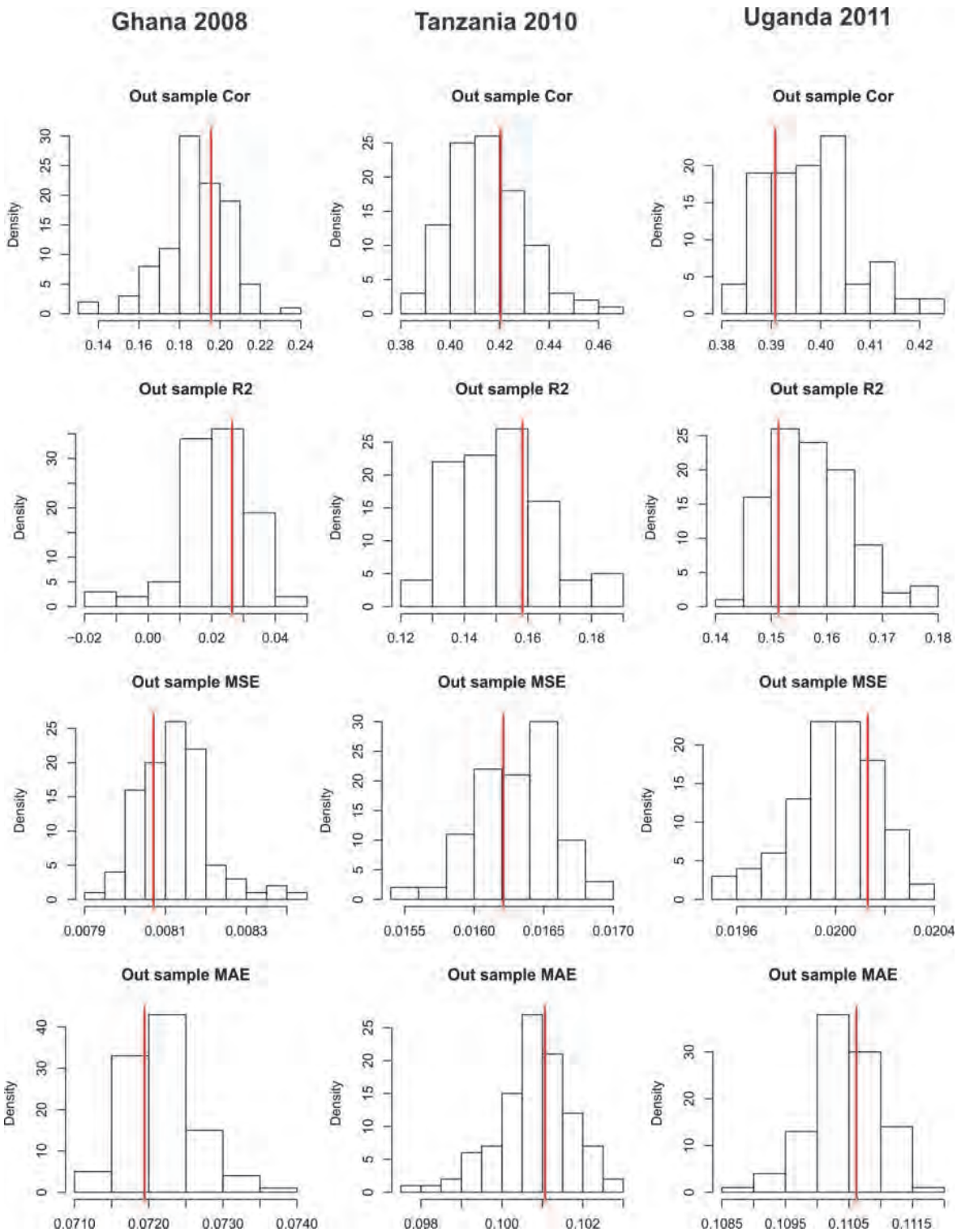


Figure 25. Effect of displacement on performance of model-based geostatistical predictive performance for the stunting indicator. Each histogram shows the distribution of one of four validation statistics (Cor = correlation between observed and predicted; R2 = predictive R^2; MSE = mean square error; MAE = mean absolute error) for models based on the 100 displaced sets relative to the non-displaced reference set (red vertical line). Panels are organized by validation statistic (rows) and DHS survey (columns).

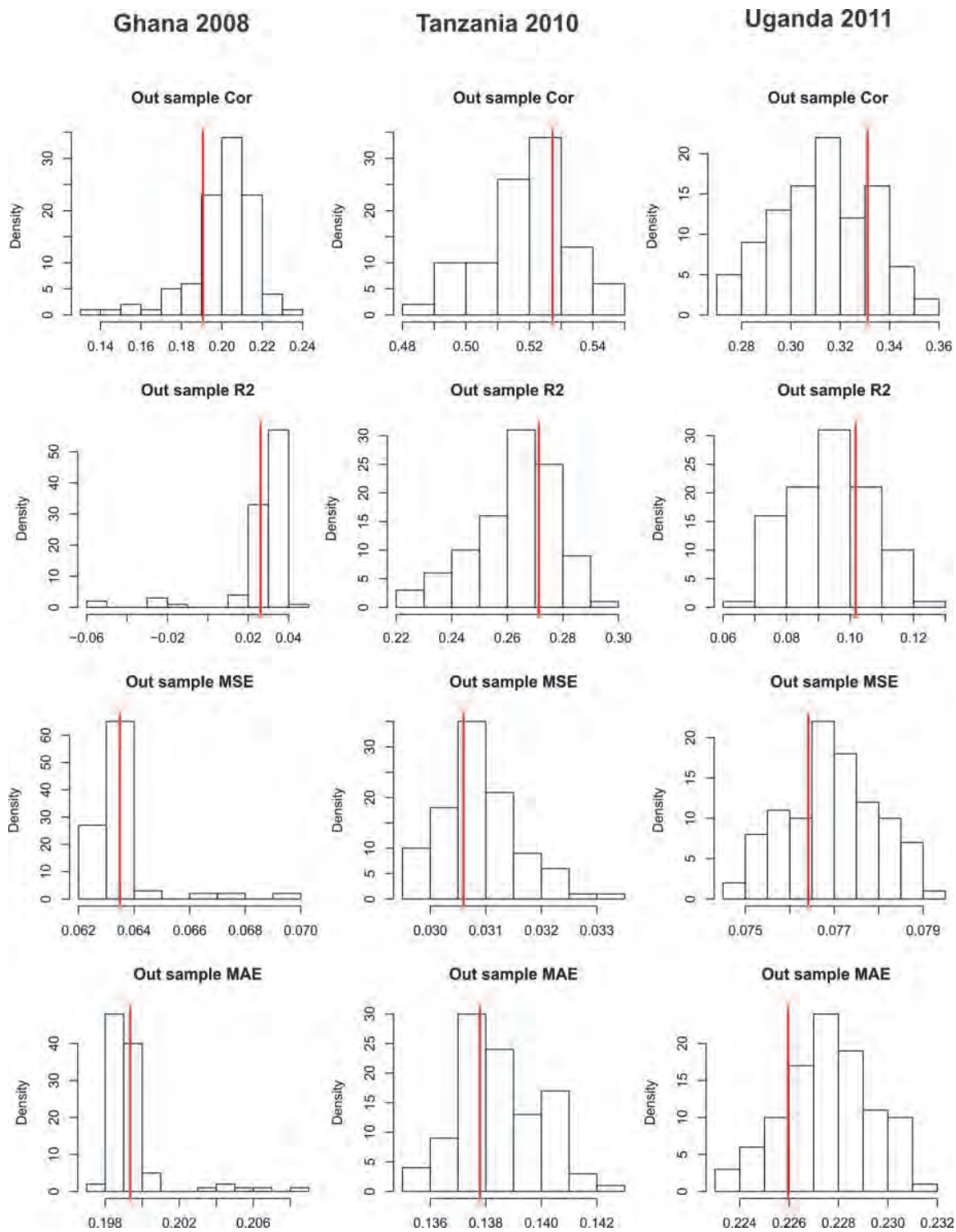


Figure 26. Effect of displacement on performance of model-based geostatistical predictive performance for the anemia indicator. Each histogram shows the distribution of one of four validation statistics (Cor = correlation between observed and predicted; R2 = predictive R^2; MSE = mean square error; MAE = mean absolute error) for models based on the 100 displaced sets relative to the non-displaced reference set (red vertical line). Panels are organized by validation statistic (rows) and DHS survey (columns).

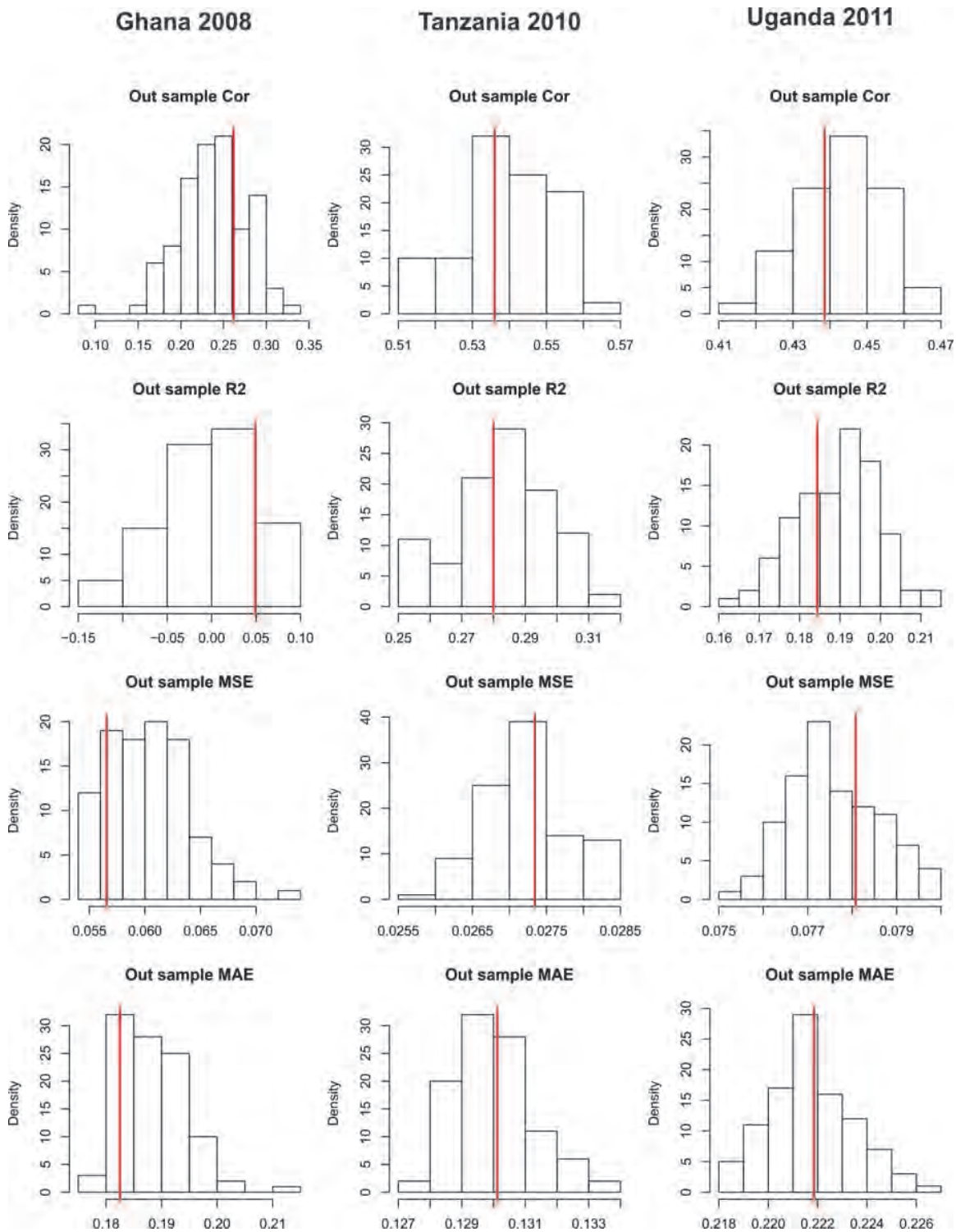
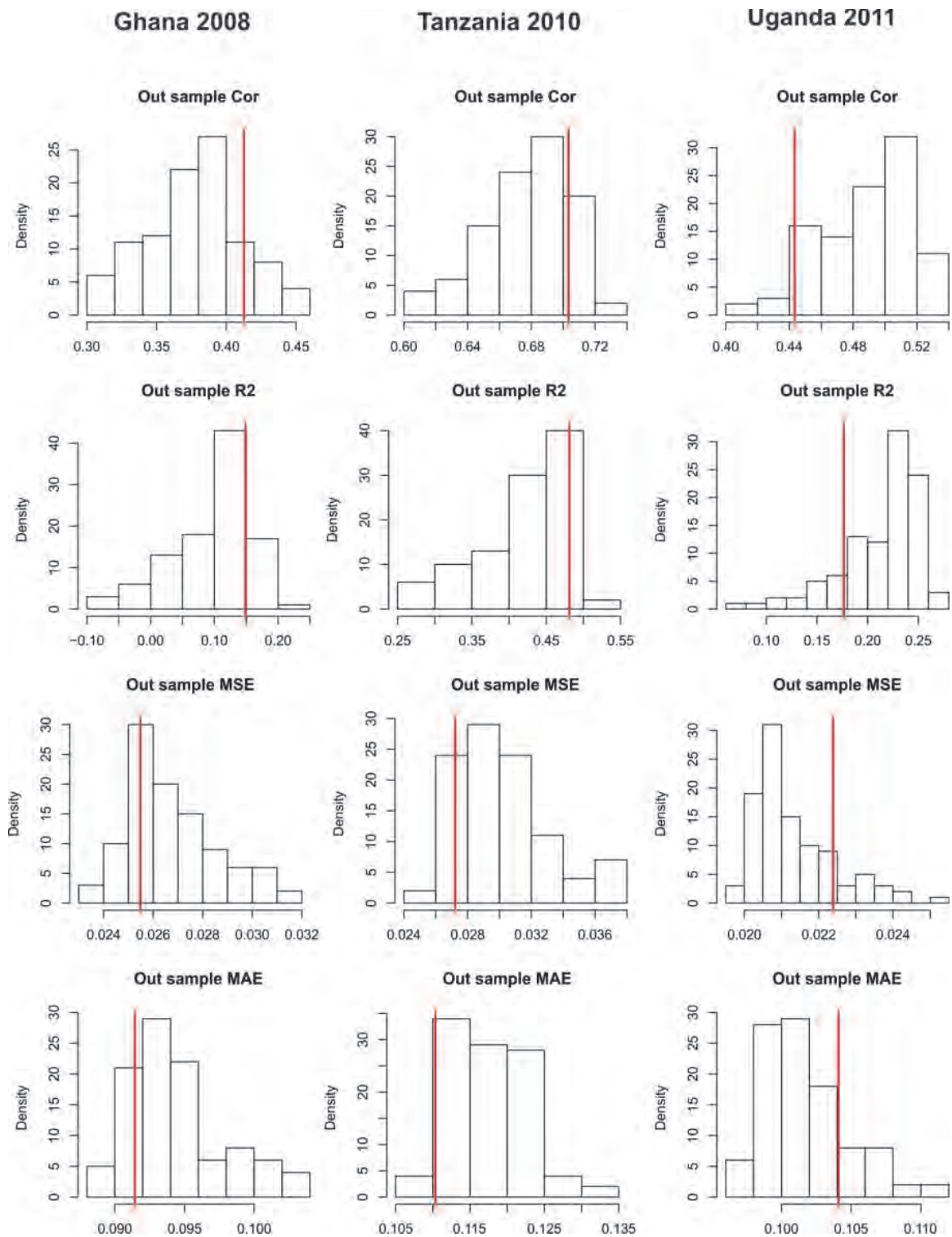


Figure 27. Effect of displacement on performance of model-based geostatistical predictive performance for the sanitation indicator. Each histogram shows the distribution of one of four validation statistics (Cor = correlation between observed and predicted; R2 = predictive R^2; MSE = mean square error; MAE = mean absolute error) for models based on the 100 displaced sets relative to the non-displaced reference set (red vertical line). Panels are organized by validation statistic (rows) and DHS survey (columns).



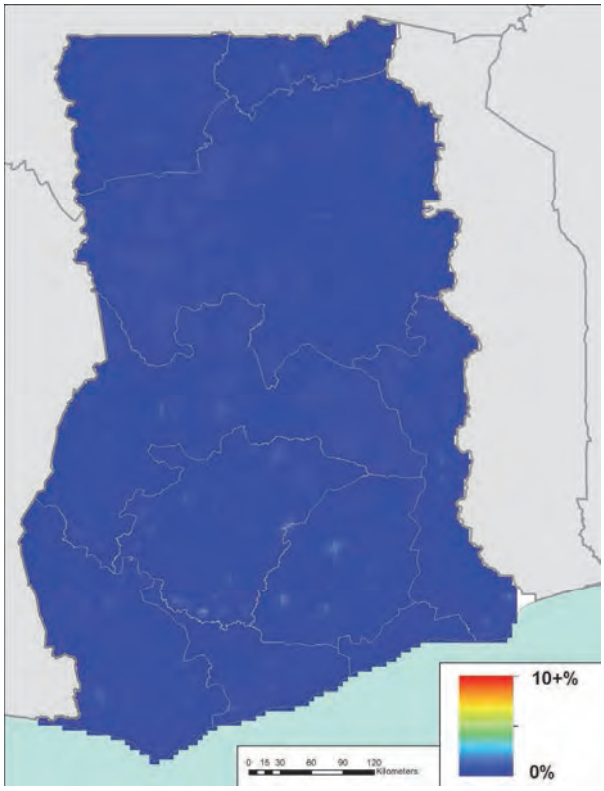
4.2.3.2 Effect on mapped surfaces

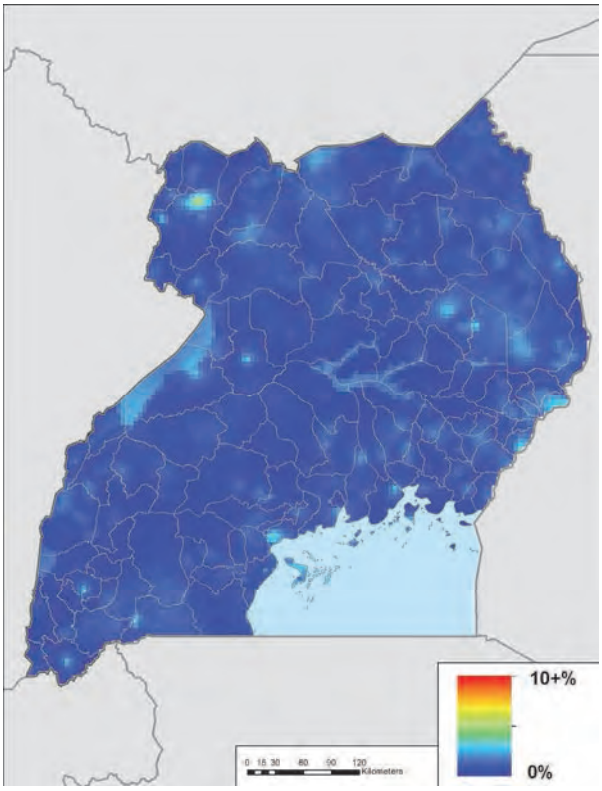
The maps shown in Figures 28 - 31 illustrate how the implications of displacement vary geographically across each interpolated surface. The value mapped in each pixel, in this case, is the standard deviation of the difference between each of the 100 predicted values (resulting from each of the 100 displacement sets) and the reference surface generated with the non-displaced data. Larger values demonstrate that the predictions for that pixel tended to differ more from the reference value; thus, the likely detrimental impact of the displacement would be larger.

Figure 28 shows results for the indicator on HIV testing for Ghana, Uganda, and Tanzania. In Ghana, where this indicator has low, relatively uniform values across the county (see original interpolated map, Figure 3), we see minimal impact of displacement for all prediction locations across the country, i.e., all 100 predictions based on the different random displacements predict a virtually identical value for the indicator at each pixel, which lead to the uniform blue map in Figure 28. The original map for Uganda (see Figure 5) was predicted with higher values but was relatively uniform across the country. Displacement effects are, in places, larger than those in Ghana, and tend to be concentrated in pockets likely associated with individual centroids that are more sensitive than others to the effects of displacement. This might be, for example, because they are located in a region where one or more underlying covariate grids are particularly spatially heterogeneous, meaning even the relatively small 2 or 5 km displacement may be enough to cause substantially different covariate values to be linked to them, which would subsequently cause markedly different predicted values in the geostatistical model. This phenomenon is even more pronounced for the Tanzania map. Scrutiny of the original interpolated surface (Figure 4) demonstrates the importance of the covariates on connectivity and access. As expected, this dependence can also be seen in the displacement map, with the more remote areas away from connecting roads displaying greater sensitivity to displacement. There are also some pronounced features near the northern border with Kenya where displacement has a large effect; this may be linked to heterogeneity in covariates in those regions.

Figure 29 shows the same set of maps for the indicator on stunting. Again, Ghana displays the least sensitivity to displacement, with the only areas of slightly raised standard deviation around the main urban areas of Accra and Kumasi. Uganda displays far more sensitivity for the HIV testing indicator, with some highly sensitive areas around the eastern and western border regions and the far south west. Tanzania displays similar patterns for this indicator than for the HIV testing one, with areas of higher sensitivity in the more remote areas and along the northern border, particularly around Mt Kilimanjaro. These patterns are similar for all three surveys for the indicators on anemia (Figure 30) and sanitation (Figure 31), although heterogeneity in Tanzania appears less related to connectivity for this indicator than to environmentally driven covariates such as elevation and temperature.

In general, across these twelve country-indicator combinations, a number of generalized conclusions can be drawn about the influence of displacement on the pixel-level predictions. First, the magnitude of difference in predictions based on displaced vs. non-displaced points is relatively modest but displays important spatial heterogeneity that, overall, can vary by at least an order of magnitude. Second, the difference tends to be larger when values of the indicators display considerable variation over short distance; when the indicator is uniform and varies little geographically, displacement can do little to disrupt accurate predictions of this “flat” response surface. Third, areas further from survey clusters tend, on average, to be the worst affected by displacement because they are more reliant on the fixed-effect (i.e., multivariate regression) component of the geostatistical model rather than drawing strength from nearby data points and this multivariate component tends to be degraded more by the dislocation induced by displacement. Fourth, there is an exception to the preceding observation where underlying covariates display marked short-scale variation, and where those covariates are important predictors (i.e., they have large coefficient values). In this case, displacement can yield very pronounced variation in the local predictions as seen around “categorical” environmental features such as mountains, rivers, and lake shore areas.





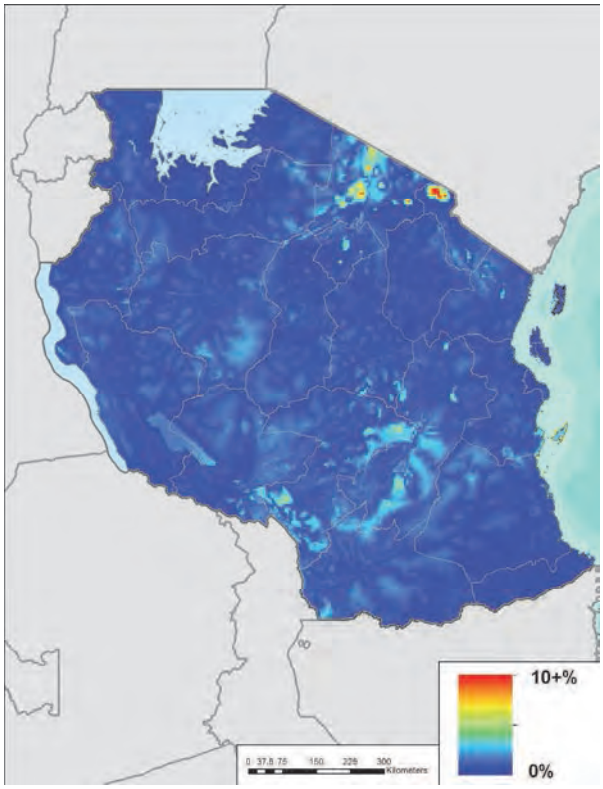
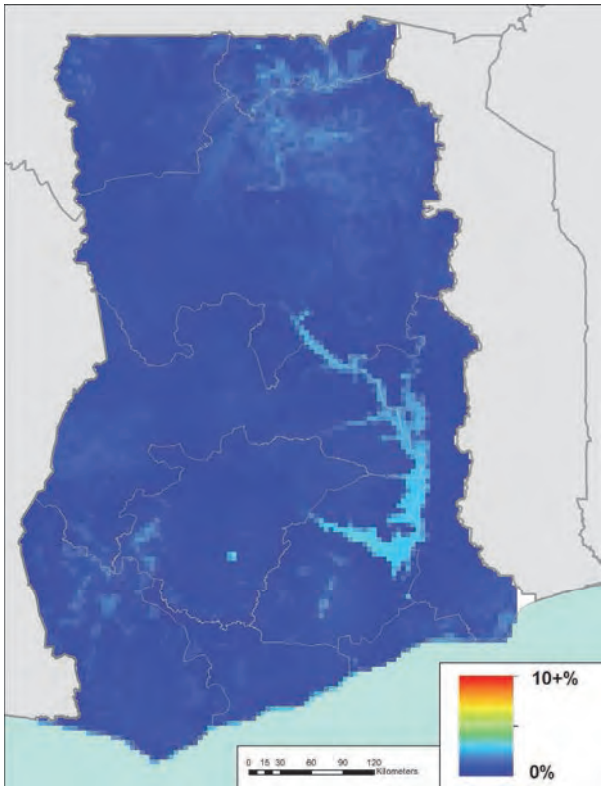
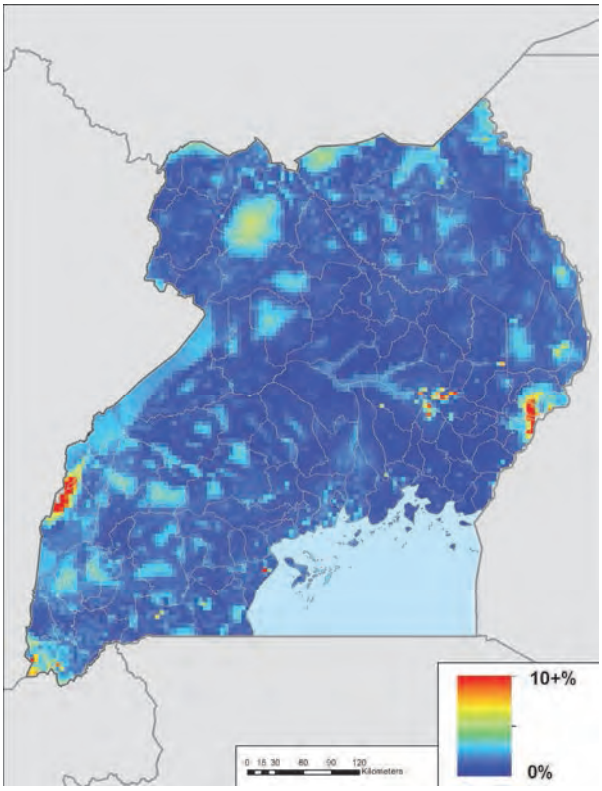


Figure 28. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for (clockwise from top-left) Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS. Maps shown are for the HIV testing indicator. The mapped variable denotes the standard deviation between 100 versions of the mapped surface each based on a different randomly displaced data set.





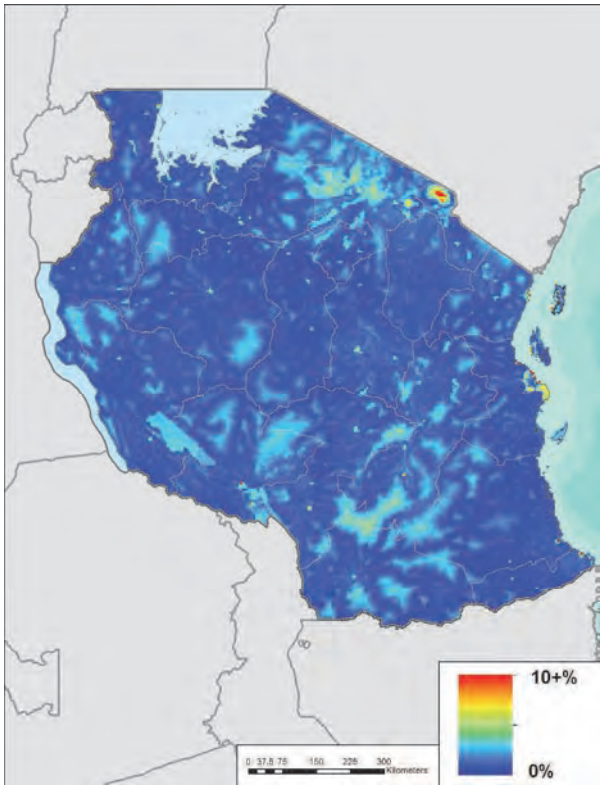


Figure 29. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for (clockwise from top-left) Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS. Maps shown are for the prevalence of stunting indicator. The mapped variable denotes the standard deviation between 100 versions of the mapped surface each based on a different randomly displaced data set.

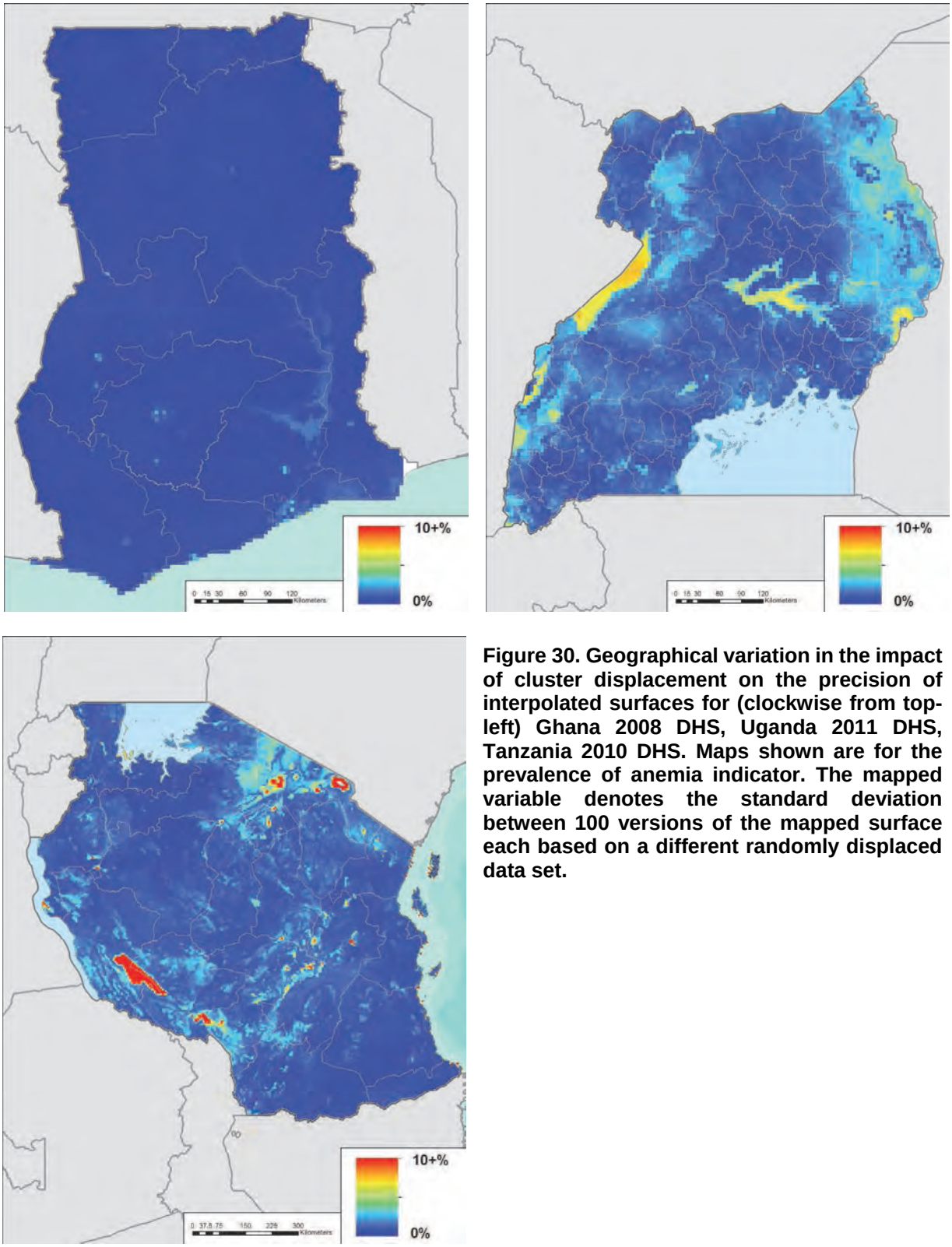


Figure 30. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for (clockwise from top-left) Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS. Maps shown are for the prevalence of anemia indicator. The mapped variable denotes the standard deviation between 100 versions of the mapped surface each based on a different randomly displaced data set.

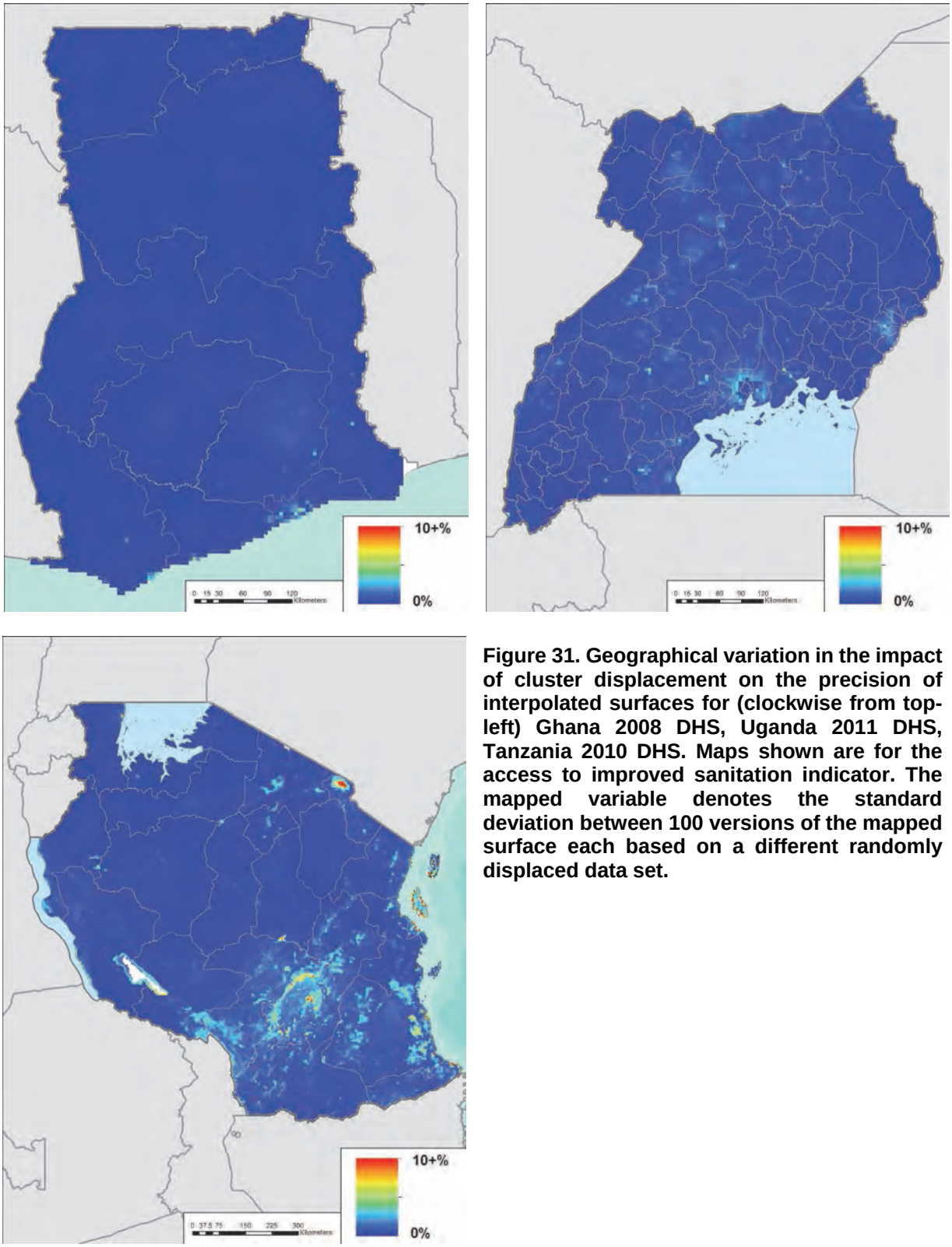


Figure 31. Geographical variation in the impact of cluster displacement on the precision of interpolated surfaces for (clockwise from top-left) Ghana 2008 DHS, Uganda 2011 DHS, Tanzania 2010 DHS. Maps shown are for the access to improved sanitation indicator. The mapped variable denotes the standard deviation between 100 versions of the mapped surface each based on a different randomly displaced data set.

5. The Potential of High Resolution Urban Mapping

5.1 Methods

5.1.1 Urban subsets of the data and urban definitions

The definition of an urban cluster can vary depending on the metrics. A set of global urban datasets was assembled to test the influence of each set on the classification of DHS clusters and whether some clusters would change an assignment from rural to urban or vice-versa by using different urban/rural maps. Each urban/rural map dataset was constructed by using a consistent satellite-based definition of "urban" or depicting metrics related to urbanicity:

- The Global Rural Urban Mapping Project (GRUMP - http://sedac.ciesin.columbia.edu/data/collection/grump-v1) produced a satellite nightlights-based dataset of urban areas at 1km spatial resolution.
- MODIS imagery was used to construct a global urban extent map at 500m resolution - http://iopscience.iop.org/1748-9326/4/4/044003
- The GlobCover global 300m resolution land cover layer (http://due.esrin.esa.int/page_globcover.php) includes an urban class.
- The Global impervious surface layer (http://www.arcgis.com/home/item.html?id=cb09f386e1694de7a778fad6e25b83a0) provides a 1km spatial resolution mapping of impervious surfaces, which is an indicator related to urbanicity.

Each dataset was assembled prior to the extraction of relevant data for Ghana, Tanzania, and Uganda.

5.1.2 High resolution urban covariates

A set of high spatial resolution (100m or finer) geospatial covariate layers for the urban areas of each of the test countries was assembled and processed. We ensured that only covariates that could be produced consistently across all DHS countries were used to demonstrate the potential of these in scaling up production of high resolution mapping in urban areas.

The Global Human Settlement Layer (GHSL - http://ghslsys.jrc.ec.europa.eu/) is a 38m spatial resolution dataset that depicts areas of human settlement globally for 2014. With such a fine spatial resolution, the variability in settlement densities could be a valuable correlate to demographic and health metrics. The final versions of the datasets are under construction; here, the alpha version was tested. The dataset was aggregated to 100m spatial resolution for urban areas in Ghana, Tanzania, and Uganda, with grid squares representing the proportion of settlement within them.

Two datasets derived from OpenStreetMap data were also constructed. OpenStreetMap is improving in terms of spatial detail and quality of data within urban areas across low income regions, especially with the support of programs such as Missing Maps (http://www.missingmaps.org/). For Uganda, Ghana, and Tanzania, data on land use and road networks were downloaded. The land use layer is inconsistent with some areas unmapped; nevertheless, it was gridded to 100m spatial resolution for use as a test covariate. The roads layer was used to construct a "distance to roads" covariate at 100m spatial resolution.

The WorldPop project (www.worldpop.org) produces 100m spatial resolution estimates of population density for low income countries across the world. The quality of the datasets is related to the resolution and year of input census data, as well as the availability of other covariate layers; however, it does provide

measures of sub-urban scale variability in population density that may be valuable for mapping population health characteristics.

Finally, for Tanzania, access was available to enumeration area level census data matched to unit boundaries for use as a covariate. This is a source of data that is increasingly becoming available in multiple countries as governments integrate geospatial technologies into census operations. For the urban areas in Tanzania, enumeration area level measures of sex ratio and proportion of the population that are economically active were calculated; these are measures that can vary substantially within urban areas, and gridded to 100m spatial resolution with areal weighting.

5.1.3 Effects of displacement on linear models in urban areas

The same cluster-level metrics of access to improved sanitation, percentage of children with anemia, and percentage of children with stunting were used to explore the effects of cluster displacement on the accuracy of linear models in urban areas in the three focus countries. These were built on the covariates outlined above. The percentage of HIV testing was not included in this analysis.

5.1.3.1 Candidate response datasets

We started with six of the raster covariates (all at 100m resolution) outlined in Section 5.2 to allow for development of resampling and plotting methods:

1. *wp* = Worldpop population density
2. *ghsl* = urban density measure from the global human settlement layer
3. *rddis* = distance to road
4. *luse* = openstreetmap land use classes
5. *sex.ratio* = sex ratio
6. *econpr* = proportion economically active

Of these, *ghsl* was transformed into a binary covariate (low (<50% of grid square containing settlement class) versus high (>50%) density), since density values were heavily left-skewed in urban areas. In addition, since *luse* classes within urban areas contained many missing data points, this dataset was subsequently saved for later consideration. For Tanzania, there were 475 clusters in the 2010 Tanzania; only 63 were located within areas classified as built-up, according to the global MODIS derived urban classification. For Ghana, there were 401 clusters, with 116 were classified as urban; for Uganda, there were 395 clusters, 95 of which were urban. We then extracted values for all 6 covariate layers at each of these cluster centroids in order to run preliminary linear models.

5.1.3.2 Buffering to account for possible bias in displaced data

One suggested way to account for possible bias due to displacement of DHS datapoints is to use buffering, wherein the covariate data are averaged within a radius around the displaced cluster centroids. We tested the influence of displacement and buffering by replicating the displacement procedure on the urban data prior to modeling. For each iteration of the resampling procedure, we did the following:

- Randomly displace each cluster centroid up to a set maximum displacement distance (as described in previous sections) and extracted new data at the displaced point.

- Ran a binomial regression model using the candidate response datasets (*wp*, *ghsl*, *rddis*, *luse*, *sex.ratio* and *econpr*) on the new dataset. In each case, we used a generalized linear model using the function glm() in R.
- Record measures of AIC, RMSE and R^2.

We performed 200 resamples at maximum displacement distances of 0.5, 1, 2.5 and 5 km. We tested for the influence of buffering on the quality of the data outputs by running the analyses on data in which the covariate values at each centroid were averaged within a buffer with a radius of 0.5, 1, 2 or 5 km. These procedures were repeated for the Tanzania, Ghana, and Uganda datasets using anemia, stunting and access to improved sanitation as response covariates. All measures of model fit and predictive power were then plotted against buffer size in each displacement category.

5.1.4 Exploration of different approaches to modeling urban settings at high resolutions

Factors that influence demographics, poverty, and health are likely to vary over fine scales in urban areas. Furthermore, the relationship between these factors and observed population-level responses in survey data may not be best predicted by using a linear regression. Therefore, we explored two methods that do not assume linear responses between covariates and data, and compared those models to binomial regression models. We then paired these models with several high-resolution covariate datasets to build the predictive models, and explored how the displacement algorithm influenced the predictive power of model results.

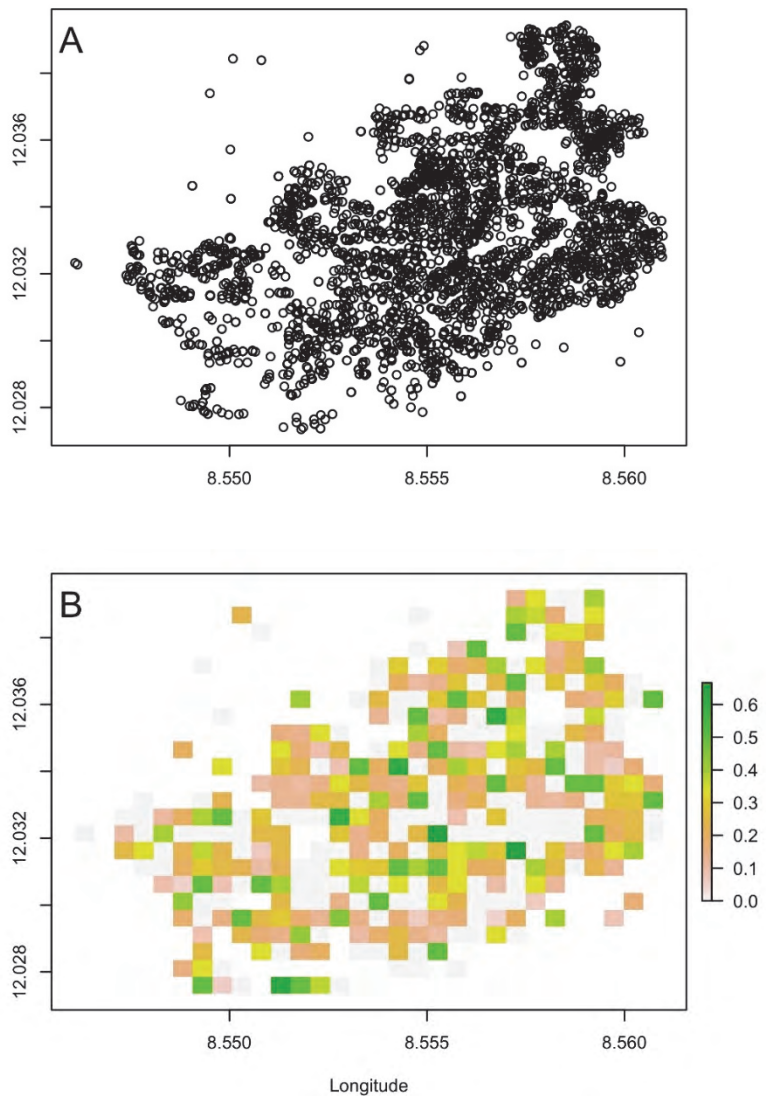
5.1.5 Kano urban dataset

The covariate layers outlined above represent those high resolution layers available today for urban areas across the low and middle income regions of the world. However, new data layers are being produced and improved, and many of these have potential for high resolution urban demographic and health indicator mapping in the future. Unfortunately, high-resolution survey data were not available within the countries originally outlined in this analysis. To explore this potential, accurate and ground-validated high resolution datasets for part of Kano City, Nigeria, assembled for population mapping efforts in northern Nigeria in collaboration with the Bill and Melinda Gates Foundation, were assembled and processed. These included the construction of the following gridded covariate datasets at 10m spatial resolution for Kano City (abbreviations are provided in brackets):

- Land use classes
- Distance to major road (dismajroad)
- Distance to tertiary road (distertroad)
- Density of roads within 100m radii (rddens100m)
- Distance to health center (dishealthcent)
- Distance to hospital (dishospital)
- Distance to market (dismarket)
- Distance to mosque (dismosque)
- Distance to school (disschool)
- Distance to waterpoint (diswaterpoint)

For validation, microcensus data of Gama ward of Kano City in Nigeria were utilized through collaboration with the Bill and Melinda Gates Foundation. All households in the ward were GPS located; all people were enumerated, and their ages were collected (Figure 32A). Experiments were designed to explore the utility of different modeling approaches. We also aimed to explore the effects of cluster displacements of varying sizes, together with the effects of different spatial resolutions of the mapping. Using the survey data, we constructed raster grids at 25, 50 and 75 m resolution and calculated the mean proportion of the population under 5 years old in each grid square (Figure 32B). We then constructed corresponding datasets of the same resolution for each of the 10 m covariate datasets. We fitted these data with the three modeling approaches described below.

Figure 32. Plots of household survey points (A) and gridded (25 m) PU5 data (B) for the microcensus surveys in the Gama ward of Kano, Nigeria.



5.1.5.1 Boosted regression trees

First, we used boosted regression trees (BRT), which use an algorithmic approach to determining which splits of the covariate data best predict the response data of interest. Because BRT do not rely on distributional assumptions, they can be used with data that are challenging for more traditional linear models.

The general approach of the BRT involves splitting the response dataset in reference to a predictor dataset, then assessing how well this split predicts the values of the response data. The splitting process is conducted a number of times for each dataset, giving rise to the concept of "trees", where each split is represented conceptually by a node that connects two divergent branches on the tree (Elith, Leathwick, and Hastie 2008). Multiple trees can be combined into one prediction to improve the predictive power of the model. Typically, determination of the effectiveness of the set of trees is made by holding back a subset of the data, and then comparing each prediction against the holdout dataset.

With larger datasets, since the number of trees available for comparison can be very large, a randomization procedure is used to limit the number of trees actually tested. In addition, it is possible to use a procedure called "boosting", in which subsequent trees are fit against the residuals of the preceding tree; this increases the predictive power from a weak model into a more powerful one. Boosting can also help avoid the problem of overfitting the data with excessive numbers of trees by monitoring the error present in the prediction for the holdout dataset. In most BRT applications, error in predictions will decline to a minimum, and then increase as the predictive model becomes overly fitted to the training data.

We used the R package "*dismo*" to build the BRTs for PU5 against the 12 high resolution covariates.

5.1.5.2 Generalized additive models

We then used generalized additive models (GAMs) to explore the use of cubic splines in predicting PU5. Briefly, GAMs fit smoothed curves for PU5 against each of the covariate data in the model, using splines as opposed to linear predictors, which are found in generalized linear models. Smoothing curves can be a range of different functions including splines, polynomials or linear terms, combined in an additive way. The GAMs have the advantage that they do not make distributional assumptions about the covariate data, although they still assume that the error in the model will follow some standard distribution. Therefore, they can be used to uncover nonlinear covariate effects relatively simply. We fit GAMs with spline-based smoothing curves on each of the covariates used, with the R package '*mgcv*'.

5.1.5.3 Linear models

To compare the efficacy of BRTs and GAMs, we explored the use of traditional linear modeling approaches, in which all covariates were modeled jointly against PU5 and the linear predictors estimated for each.

5.1.5.4 Effects of displacement

Finally, we performed a series of displacement resampling with maximum displacement values between 0 and 2 km at 500m intervals. We did not run analyses on buffered data since the aim was to investigate the use of high-resolution data, and the rasterizing process resulted in an average value within each grid cell. For the purposes of comparing the predictive power and error in each model, we held back a random subset of 20% of the data set for cross-validation and used these data to calculate correlation coefficients (PR^2) and RMSE.

5.2 Results

5.2.1 Urban definitions

For all cluster centroids in the Tanzania, Ghana, and Uganda datasets, population density estimates were extracted from the WorldPop project (www.worldpop.org) population density 2010 datasets for the centroid of each cluster. This enabled us to assess the range of population densities covered by the "urban" and "rural" defined clusters and to examine any outliers (e.g., low population density "urban" clusters or high population density "rural" clusters). We also extracted the urban/rural assignment of each cluster centroid based on the satellite-based urban classifications described above; this enabled examination of levels of mismatch between DHS-defined urban/rural cluster classification, and the satellite-based classifications.

Figure 33 and Figure 34 show plots where each DHS cluster is represented by a bar, with the height of the bar representing the estimated population density, and the bars ordered by these density estimates. The color of the bar indicates urban/rural assignment of each cluster by the different classifications.

Figure 33 shows that, as anticipated, the vast majority of clusters classified as urban in the DHS surveys have much higher estimated population densities than those classified as rural. There are outliers, particularly in Ghana, where there are many urban clusters in areas estimated to have low population densities. The three satellite-derived urban classifications generally show a clearer relationship with population density for all three countries, with clear divides in terms of population densities between clusters defined as urban and rural by the satellite metrics.

Figure 33. Clusters represented by bars, with heights proportional to log(population density) and ordered by population density. The bar for each cluster is colored by whether it is classified as urban or rural according to the different indicators listed on the right-hand y-axis.

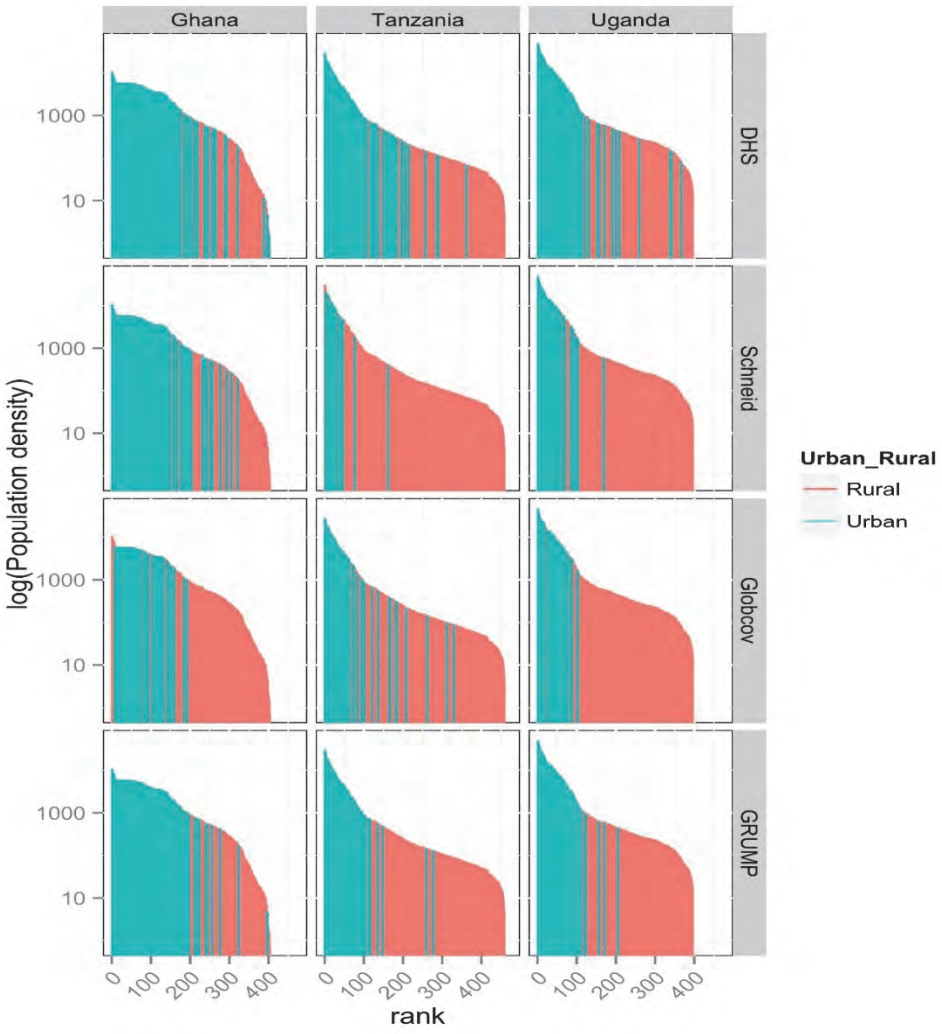
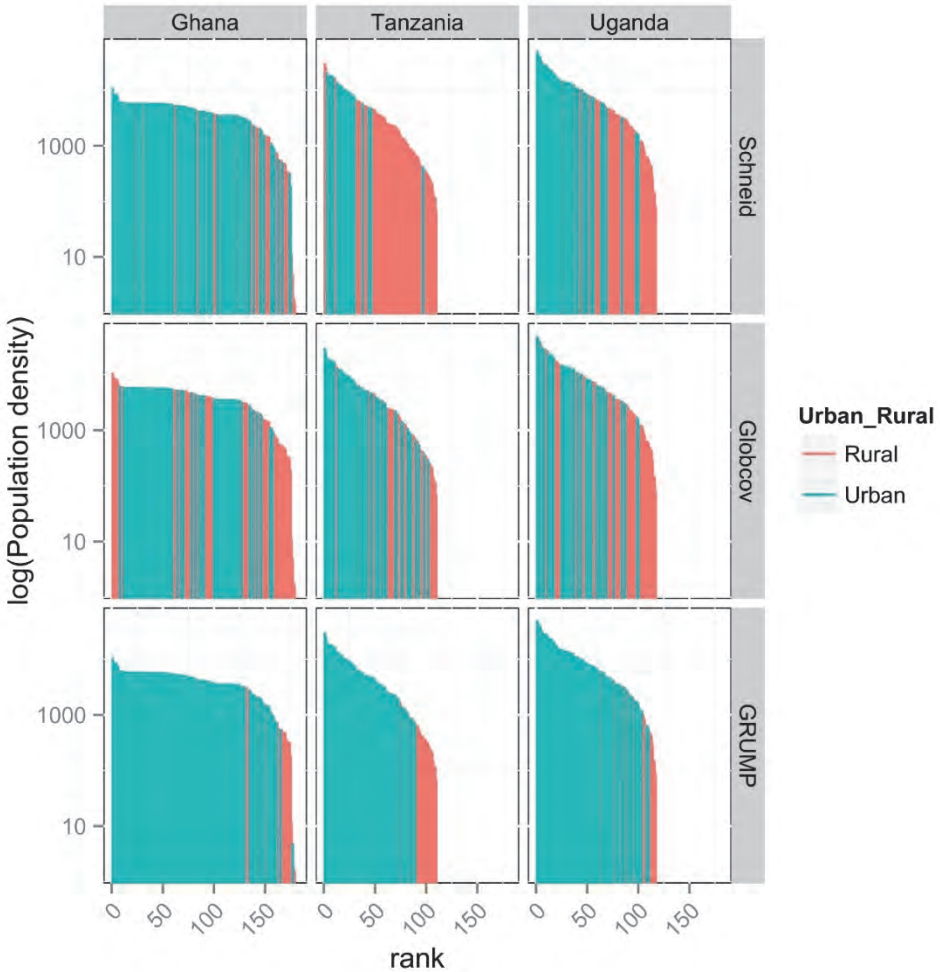


Figure 34, which plots only the clusters defined as urban in the DHS surveys, shows how consistent these classifications are with the satellite derived measures. Generally, there is good agreement between indicators with the large majority of DHS-defined urban clusters also classified as urban by the satellite metrics. However, it is clear that mismatches occur, especially against the Schneider et al MODIS definitions for Tanzania, where most of the DHS defined urban clusters are classed as rural by the MODIS-based definition.

Figure 34. Urban clusters as defined by the DHS classification represented by bars, with heights proportional to log(population density) and ordered by population density rank. Each bar is colored by the urban/rural assignment of different satellite-derived urban/rural maps.



5.2.2 Effects of displacement on model estimates

This section summarizes the results of the displacement exercise in urban areas. There are several general patterns in our resampling results. In general, models in urban areas have much lower R^2 than country-wide models. This is due to a combination of factors such as the smaller sample sizes and reduced covariate set used in the models.

Another result showed that the variability in model diagnostic values was lower at higher buffer values; this suggested that buffering was smoothing out some of the variability due to random selection of data points within the displacement area. The RMSE values also tend to increase as buffer size increases. There is also a slight but negligible increase in RMSE values within increased maximum displacement distance. When buffer size increases, R^2 values tend to increase or remain the same, although there is negligible difference in R^2 with increased displacement distance. Finally, AIC values remain the same with increased buffer size, but are smaller with increasing displacement distance. An exception to these general patterns occurs when the displacement distance is set to 5 km, when we see large variation in all measures of model fit, marked decreases in RMSE, and increases in R^2. Country-specific results from models for anemia, stunting, and access to sanitation are discussed below.

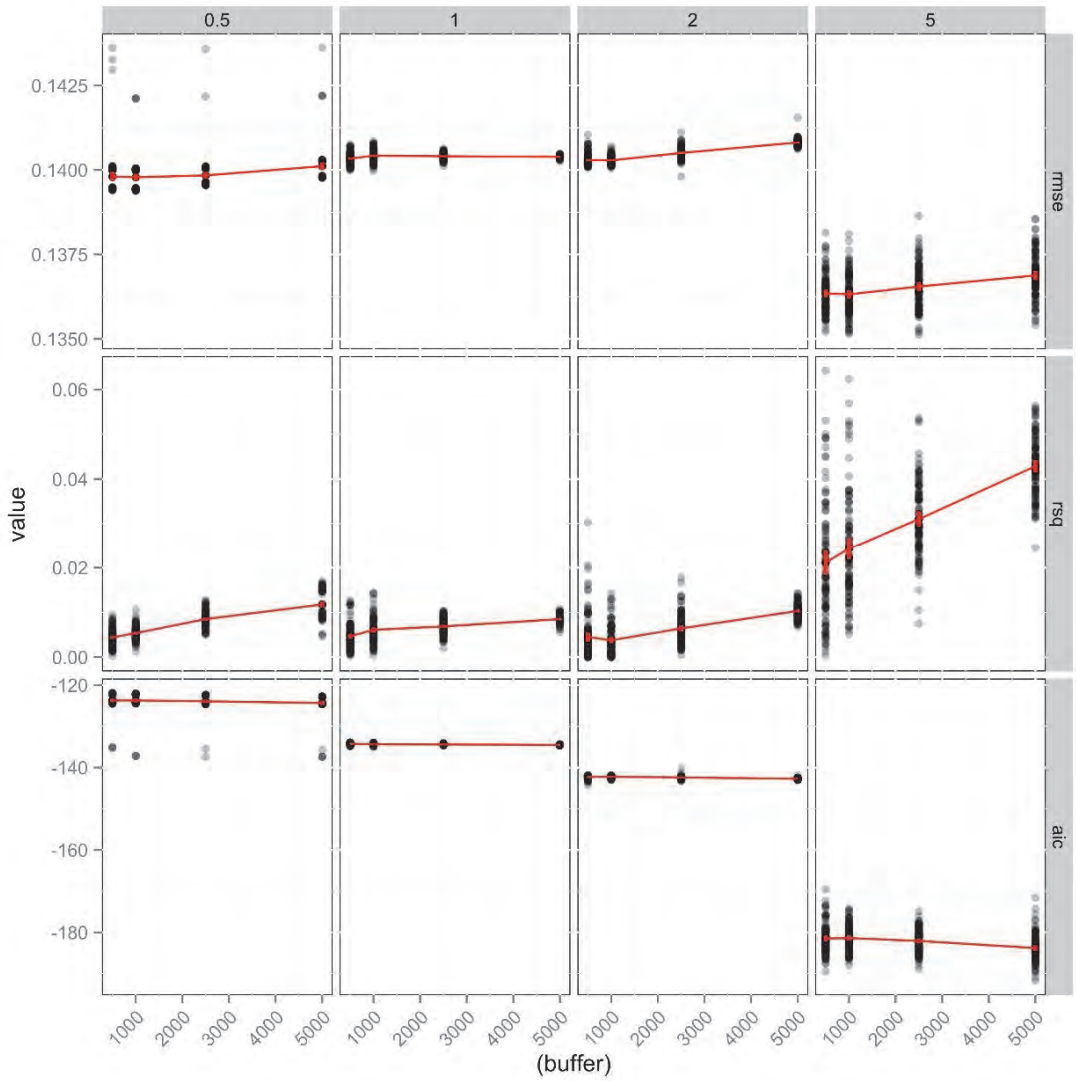
5.2.2.1 Stunting in children

All models had low R^2 values, with generalized linear models explaining between 1% and 8% of the variance in the data, depending on the country and levels of displacement and buffering.

Tanzania

Model results for stunting in children suggest that the selected covariates had very little explanatory power, except when displacement approached 5 km (Figure 35). There was little variation in R^2, RMSE or AIC with increased buffer size. Measures of model fit and bias were best when the buffer is 5 km (D= 5000). However, as discussed previously, this may be an artifact of the large degree of displacement possible at this level.

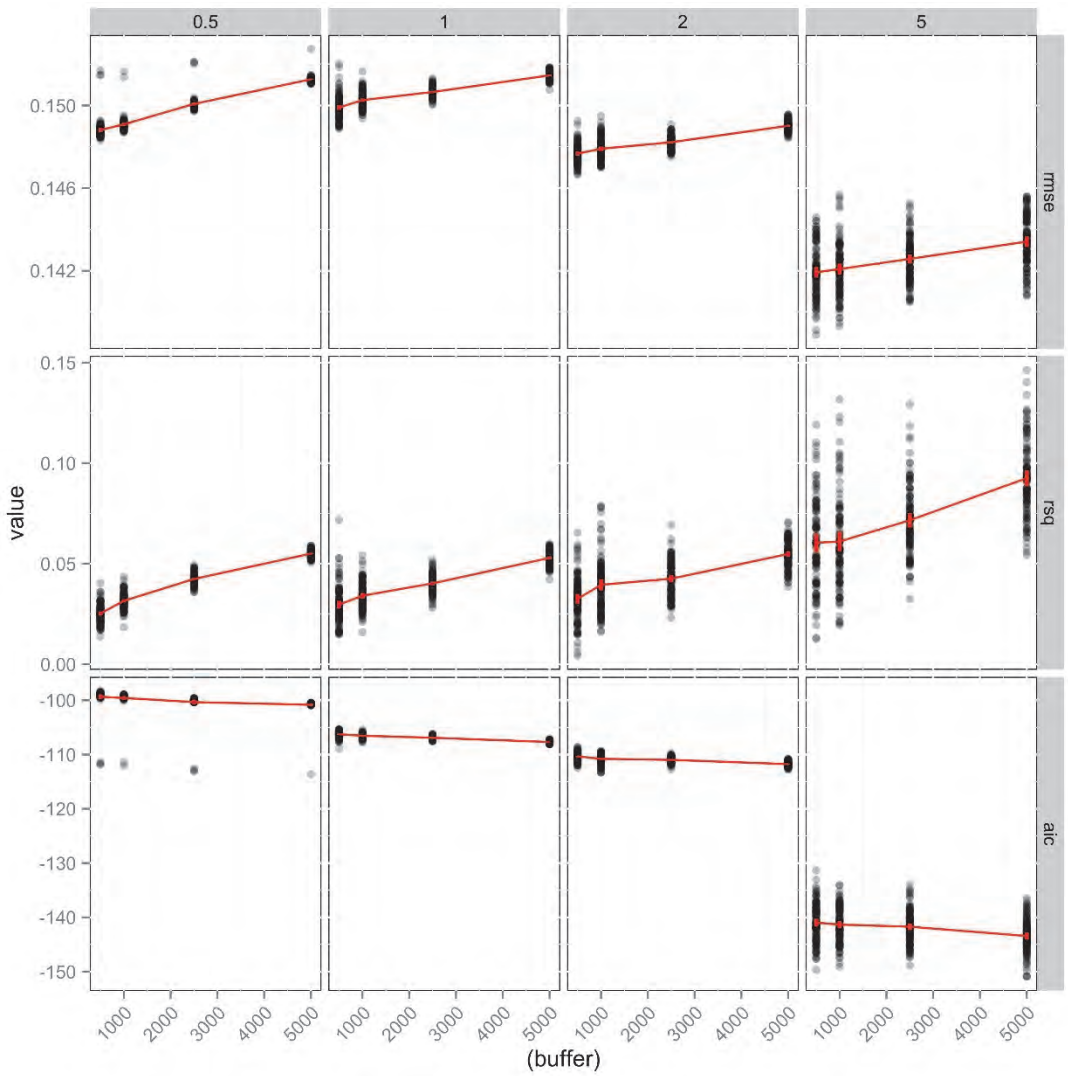
Figure 35. Effects of displacement and buffering on diagnostics for models on the prevalence of stunting in children in Tanzania. Each point represents the diagnostic value (RMSE, R^2 or AIC) for one iteration of the displacement procedure. Columns indicate different levels of maximum displacement (in km), while rows provide different model diagnostics.



Uganda

In Uganda, R^2 values were low (0.025) at small buffer sizes, but increased to nearly 0.07 at large buffers (Figure 36). There was no difference in R^2 with increasing buffer distance, except at D = 5 km. The RMSE values dropped slightly between D = 0.5 and 2 km, then dropped significantly at D = 5 km. The AIC values also dropped with increased D.

Figure 36. Effects of displacement and buffering on diagnostics for models on the prevalence of stunting in children in Uganda. Each point represents the diagnostic value (rmse, R^2 or aic) for one iteration of the displacement procedure. Columns indicate different levels of maximum displacement (in km), while rows provide different model diagnostics



Ghana

In Ghana, our models had essentially no predictive power, with R^2 varying between 0 and 0.01. Consequently, these results were omitted from the analysis.

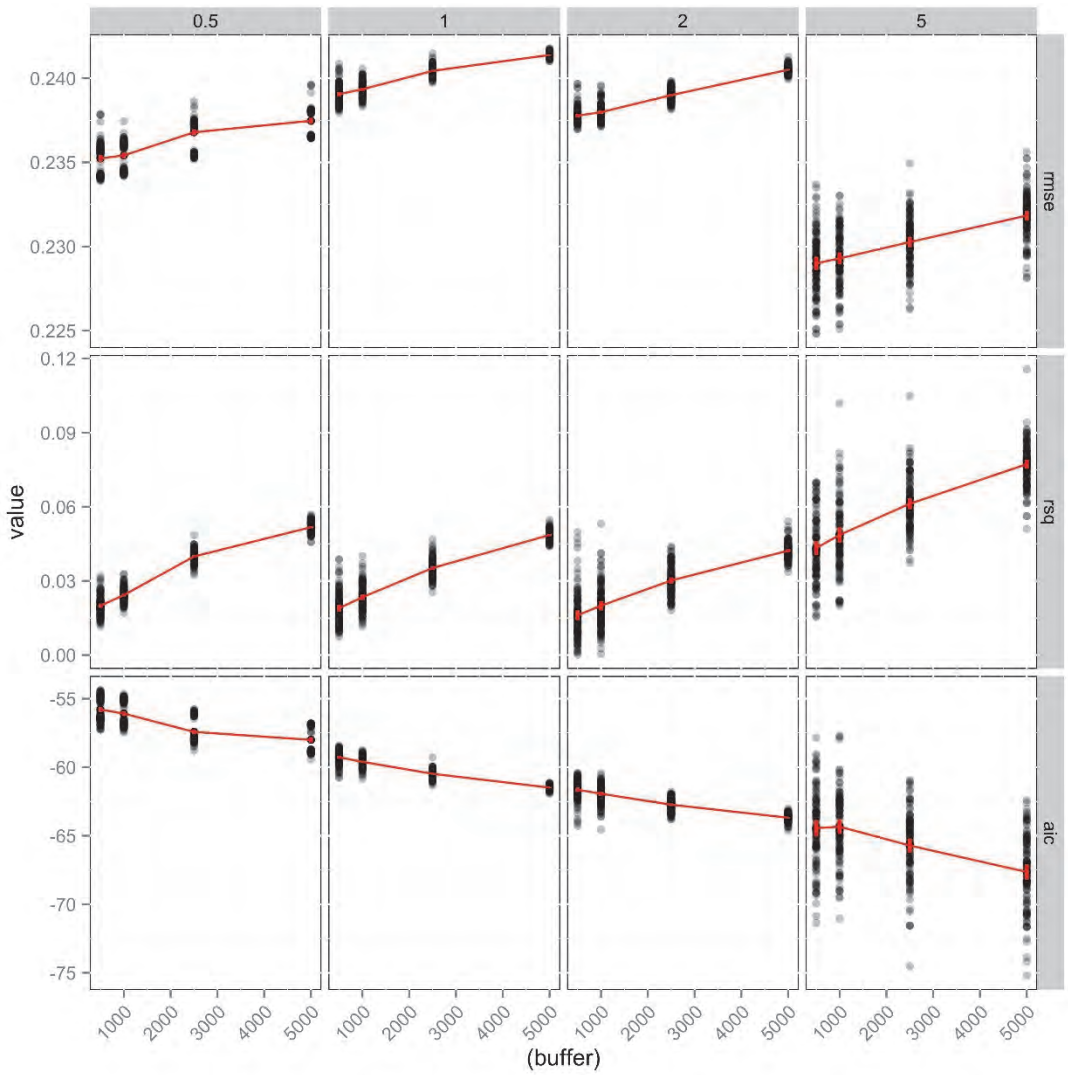
5.2.2.2 Anemia prevalence in children

Models for anemia prevalence had relatively higher R^2 values for Uganda and Tanzania, but not for Ghana. However, there were strong differences in how these values changed with varying buffer and displacement size.

Tanzania

In Tanzania, models for anemia had lower R^2 values at smaller buffer sizes across all values of D, increasing from a low of 0.03 to a high of ~0.05 (at D=2) (Figure 37). The RMSE values increased slightly between D=0.5 km and D=2 km, then dropped at D=5 km, while AIC values dropped consistently with increased D and increased buffer.

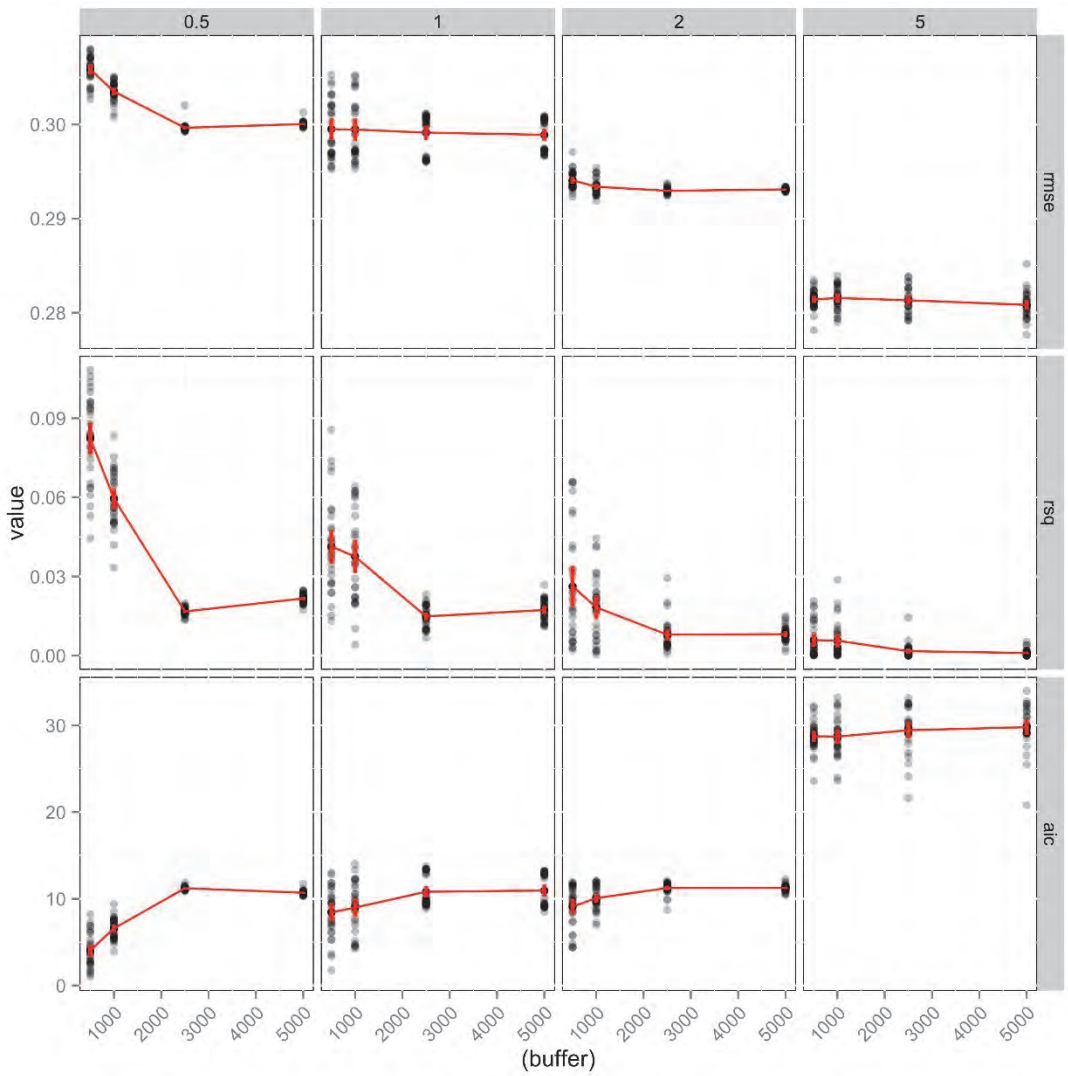
Figure 37. Effects of displacement and buffering on diagnostics for models on the prevalence of anemia in children in Tanzania. Each point represents the diagnostic value (rmse, R^2 or aic) for one iteration of the displacement procedure. Columns indicate different levels of maximum displacement (in km), while rows provide different model diagnostics.



Uganda

Models for anemia in Uganda showed a markedly different pattern to those in Tanzania (Figure 38). Here, R^2 values were highest for low displacement and low buffer sizes, and then dropped from a high of almost 0.09 to a low of ~0.01. This suggested a strong degradation in the correlation of response data with prediction data. This trend is also reflected in the AIC values, which increased with increasing buffer and D.

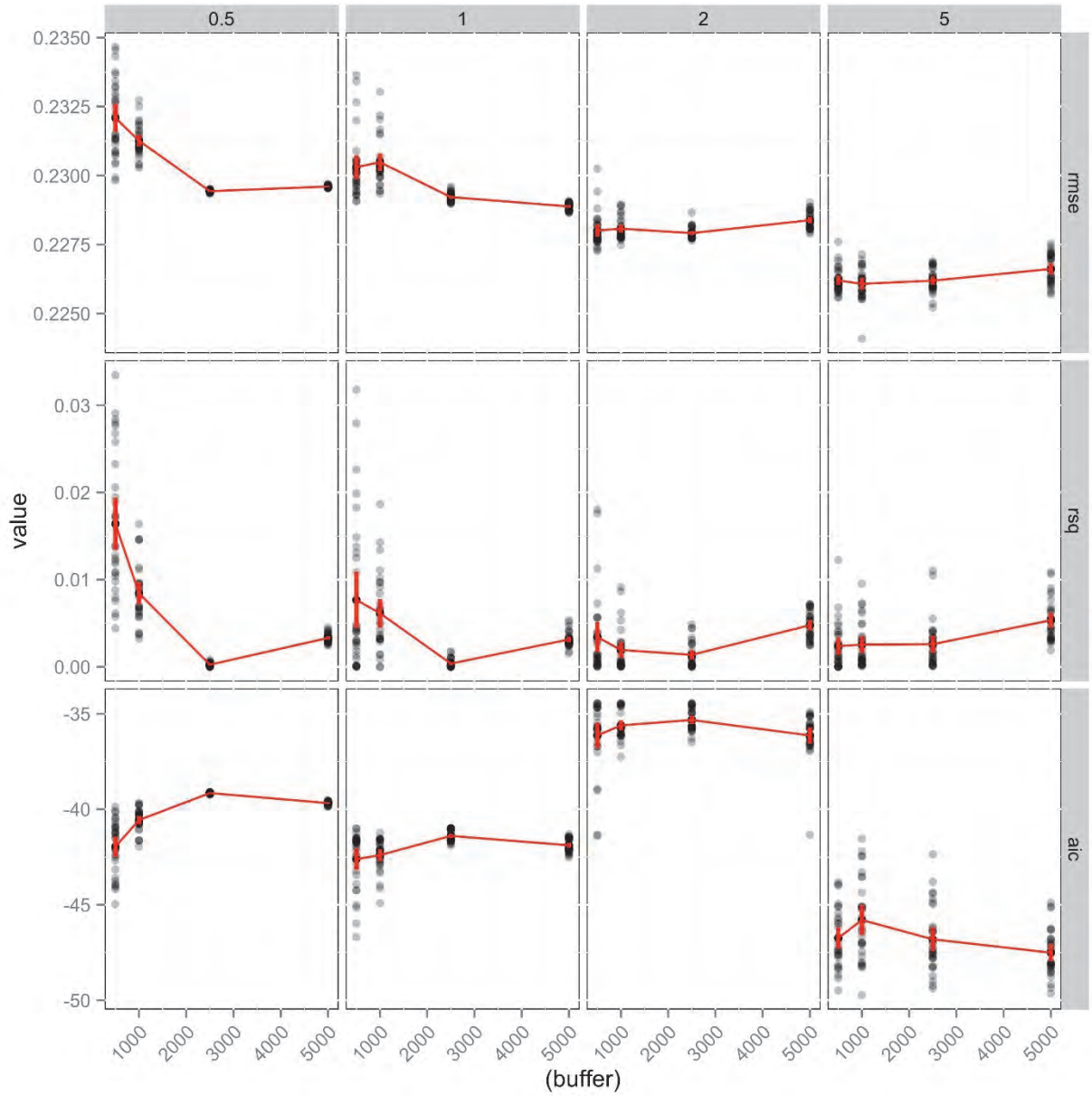
Figure 38. Effects of displacement and buffering on diagnostics for models on the prevalence of anemia in children in Tanzania. Each point represents the diagnostic value (rmse, R^2 or aic) for one iteration of the displacement procedure. Columns indicate different levels of maximum displacement (in km), while rows provide different model diagnostics.



Ghana

In Ghana, linear models had very low power to predict anemia, with R^2 values varying between 0.01 and 0.03 (Figure 39). Maximum predictive power was achieved with the lowest displacement levels and lowest buffer values.

Figure 39. Effects of displacement and buffering on diagnostics for models on the prevalence of anemia in children in Ghana. Each point represents the diagnostic value (rmse, R^2 or aic) for one iteration of the displacement procedure. Columns indicate different levels of maximum displacement (in km), while rows provide different model diagnostics.

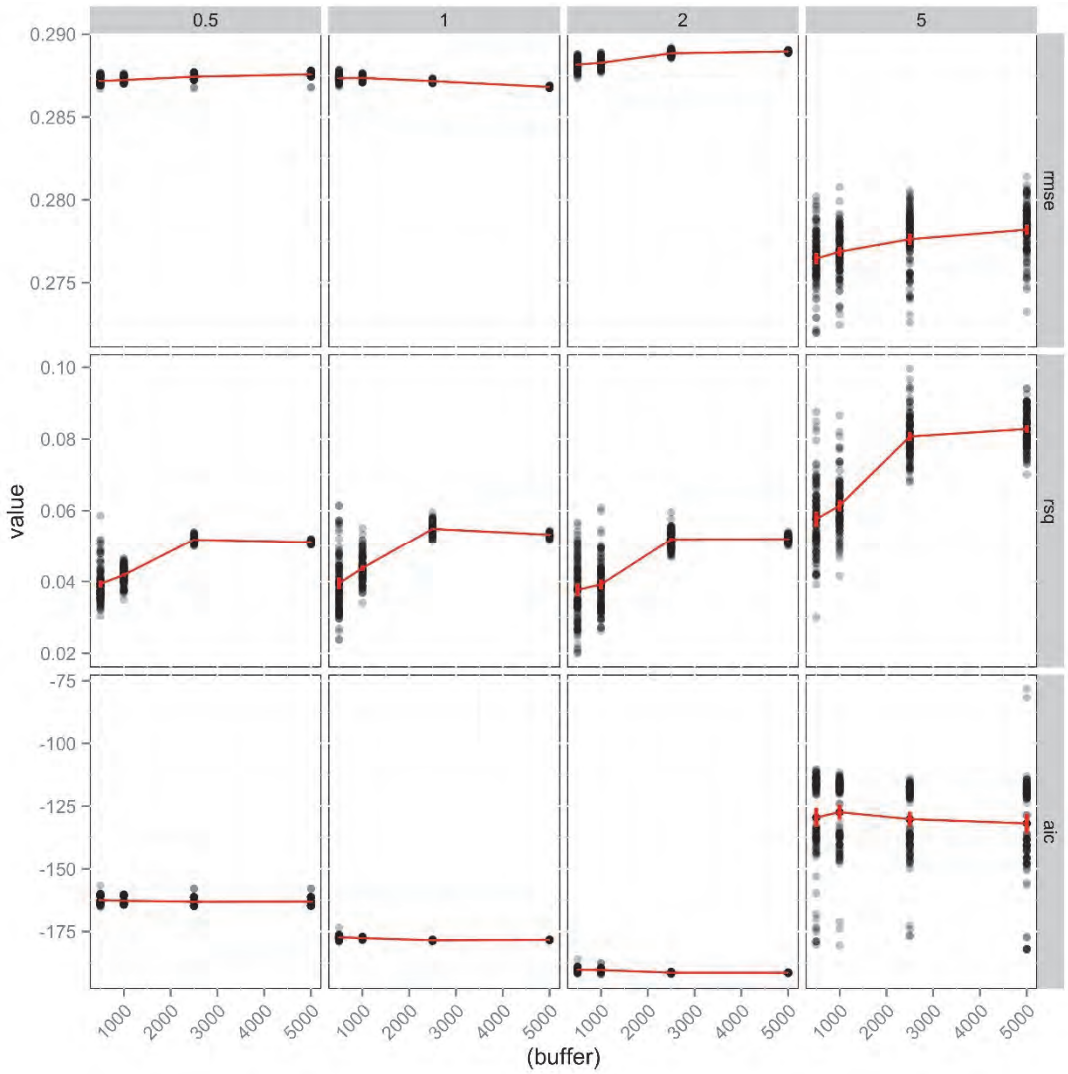


5.2.2.3 Access to improved sanitation

Tanzania

In Tanzanian models for improved sanitation (Figure 40), there was a general increase in R^2 with increasing buffer sizes and stable R^2 with increase displacement, except at D=5 km. The RMSE values were stable across all buffer sizes and displacements, except D=5 km. The AIC values dropped between D=0.5 and D=2 km, then rose at D= 5km.

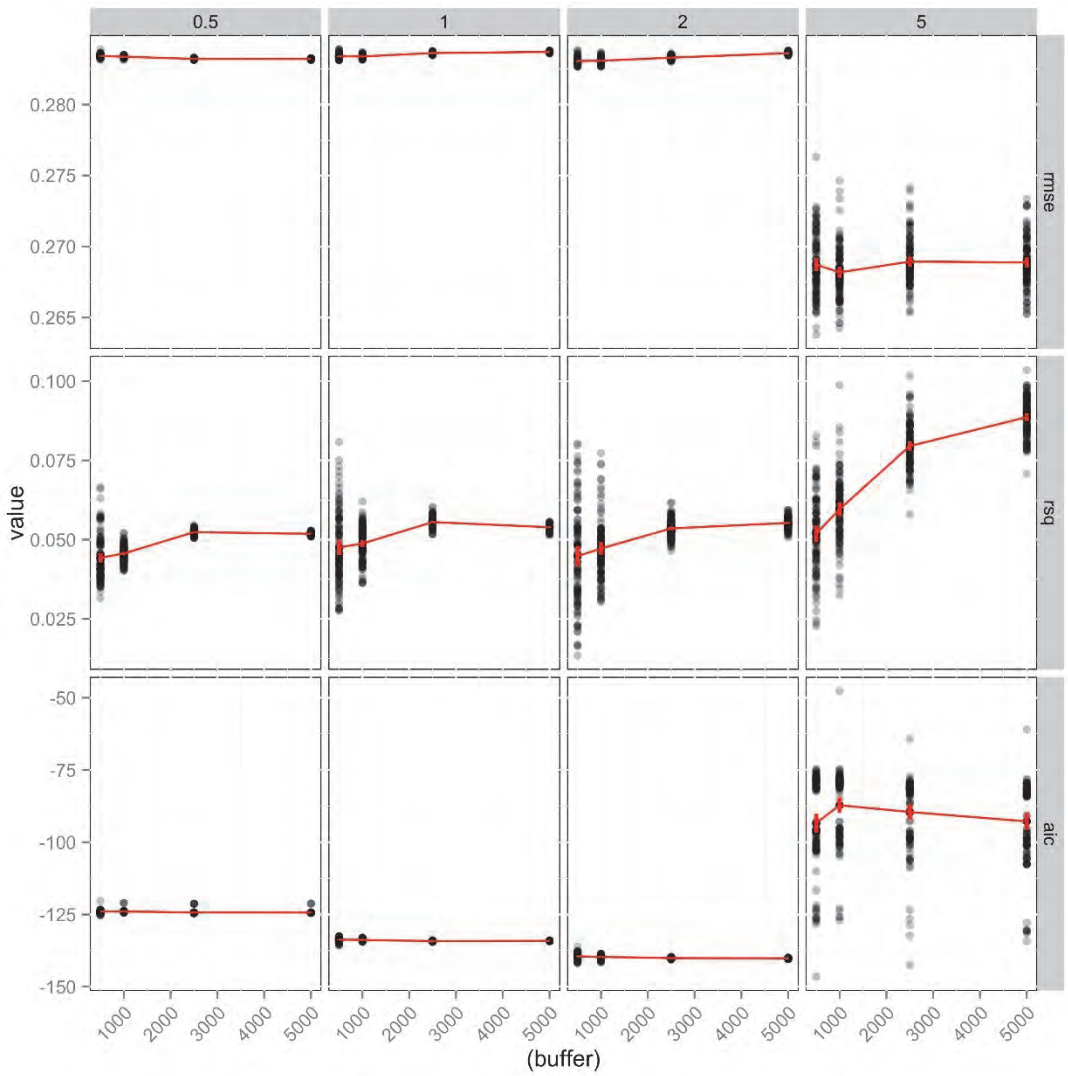
Figure 40. Effects of displacement and buffering on diagnostics for models on the prevalence of improved sanitation in Tanzania. Each point represents the diagnostic value (rmse, R^2 or aic) for one iteration of the displacement procedure. Columns indicate different levels of maximum displacement (in km), while rows provide different model diagnostics



Uganda

Models for improved sanitation in Uganda (Figure 41) showed a similar pattern to Tanzania, with R^2 increasing with increasing buffer sizes but relatively stable R^2 with increasing displacement, for D≤2 km. The RMSE values were stable across all buffer sizes and displacements, with the exception of D=5 km. The AIC were lowest at D=2 km, then rose at D=5km.

Figure 41. Effects of displacement and buffering on diagnostics for models on the prevalence of improved sanitation in Uganda. Each point represents the diagnostic value (rmse, R^2 or aic) for one iteration of the displacement procedure. Columns indicate different levels of maximum displacement (in km), while rows provide different model diagnostics



Ghana

Data on improved sanitation in Ghana were unsuitable for linear modeling, with only one non-zero value in Urban areas. Consequently, these analyses were omitted from this section.

5.3 Exploration of Model Approaches to High-resolution Datasets

5.3.1 Comparison of Generalized Additive Models (GAM), Boosted Regression Trees (BRT) and Linear Models (LM) for predicting proportion of children under 5 years old

Here, we compare the outputs for the linear, BRT and GAM models for household-level data. The household level linear models had lowest predictive power, with $PR^2 = 0.145$. Values for LM ($PR^{2=}0.23$) and BRT ($PR^{2=}0.25$) were higher. All models had similar RMSE values that were close to 0.171.

However, each model placed differing levels of importance on each of the covariates that were used to predict the proportion of children under 5. In the BRT model, distance to market, distance to waterpoint, health centers and schools, as well as road density, longitude and latitude were all relatively important (Table 10). For LMs, latitude, distance to small roads, distance to mosque, distance to medical center, distance to market, and distance to health center all had significant p-values (Table 11). Finally, for the GAM model, distances to major roads, schools, health centers, markets, tertiary roads, and density of tertiary roads were all relatively important (Table 12).

It is difficult to compare the relative importance of covariates in each of the models, given the different methodologies used for each and the fact that some of these covariates may have been closely related to one another. However, in general, factors such as distance to school, distance to health (or medical) center, distance to market, and distance to roads all appeared to be more important in all three models.

Table 10. Summary of relative influence for different covariates in the boosted regression tree model for proportion of children under 5. *Rel.inf* is a measure of the relative influence of each covariate in explaining variation in the response data.

Variable	rel.inf
Distance to market	24.74
Distance to water point	18.24
Longitude	8.59
Distance to Medical Centre	8.09
Density of roads (1 km)	7.32
Latitude	7.05
Distance to school	6.33
Distance to main roads	4.85
Distance to health centre	4.23
Distance to Mosque	3.77
Density of roads (100 m)	3.52
Distance to small roads	2.77
Distance to tertiary roads	0.49
Distance to hospital	0.00

Table 11. Summary of the parameters for different covariates in the full linear model (LM). *Estimate* indicates the parameter value estimated in the model, while *Std.Error* provides a measures of uncertainty in the estimate. P-value indicates the significance of the parameter estimate, and with low p-values <0.05 are highlighted in bold. Covariate names are codified, with *rddens* indicating density of roads within a certain radius, *dis* indicates minimum distance from various features such as schools, roads mosques etc.

	Estimate	Std. Error	p-value
(Intercept)	-296.9669	97.9571	**0.0024**
Distance to main road	-16.172	3.1091	**0**
Distance to health centre	-26.1521	7.4797	**0.0005**
Distance to mosque	-10.2308	3.4514	**0.003**
Latitude	22.7357	7.7502	**0.0034**
Distance to small roads	-39.5252	15.9561	**0.0133**
Distance to medical	13.3137	6.1392	**0.0302**
Distance to market	6.7493	4.0294	0.094
Distance to school	-7.0388	5.3641	0.1895
Road density (100m)	0	0	0.2393
Distance to water point	-5.4094	8.9008	0.5434
Longitude	2.7519	6.0439	0.6489
Distance to tertiary road	3.8384	9.5441	0.6876
Road density 1(km)	0	0.0001	0.8128

Table 12. Summary of the relative influence of different covariates in the Generalized Additive Model at the household level. *Edf* indicates effective degrees of freedom, and provides a measure of the number of inflection basis points used in fitting the spline. Where edf=1, the term is the same as a straight line. *Ref.df* is the residual degrees of freedom, *F* is the F statistic for each term and *p.value* indicates the significance of each. The notation s() indicates the cubic splie function used on each covariate.

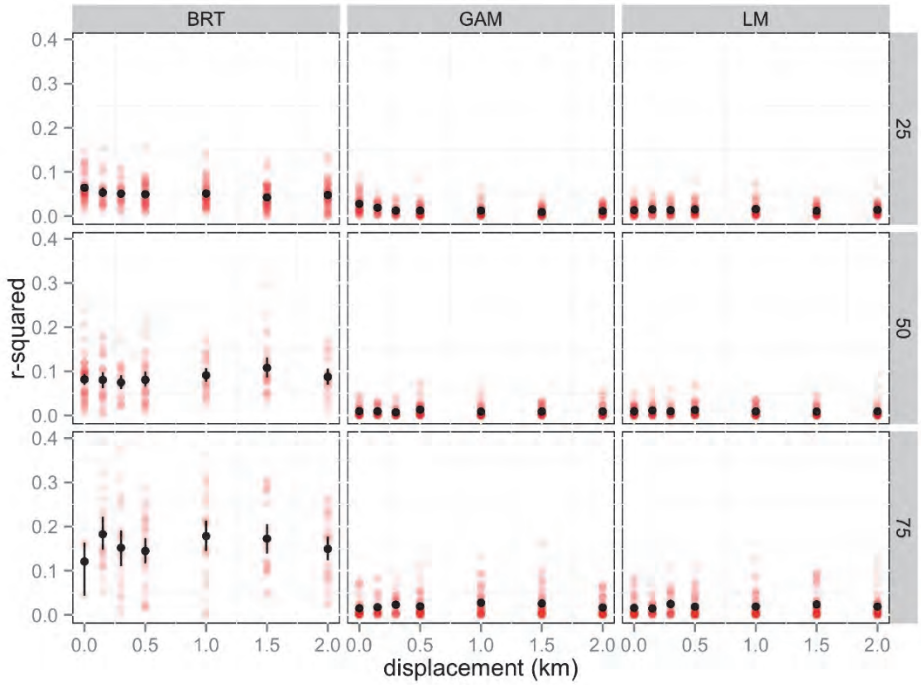
	edf	Ref.df	F	p.value
s(Distance to major road)	1.00	1.00	11.13	**0.00**
s(Distance to school)	8.26	8.79	2.96	**0.00**
s(Distance to health centre)	1.10	1.17	7.23	**0.01**
s(Distance to market)	8.43	8.89	2.33	**0.01**
s(Distance to small road)	1.00	1.00	3.80	**0.05**
s(Road density (100 m))	4.25	5.36	2.14	**0.05**
s(Road density (1 km))	5.13	6.43	1.94	0.07
s(Distance to mosques)	1.00	1.00	3.12	0.08
s(Distance to waterpoints)	8.12	8.67	1.70	0.09
s(Latitude)	5.47	6.80	1.11	0.35
s(Longitude)	4.76	6.03	0.78	0.59
s(Distance to tertiary roads)	1.00	1.00	0.19	0.66
s(Distance to hospital)	3.79	4.23	0.11	0.98
s(Distance to medical centre)	3.79	4.23	0.11	0.98

5.3.2 Effects of displacement on models at differing resolutions

5.3.2.1 Proportion of the population that are children under 5 years old

In our analyses of the high-resolution Kano urban sampling data, the displacement analysis reveals several patterns in the relative predictive power of the three different models and the influence of displacements on model predictions in urban areas (Figure 42). In general, we found that Generalised Additive Models (GAM), Boosted Regression Trees (BRT) and Linear Models (LM) all had lower predictive power with aggregated data than with household data. In LM at all displacements and data resolutions, PR^2 was near to 0. For GAMs, we found that in the highest resolution data (25 m grid cells), PR^2 was highest (~0.03) at 0 km displacement, and then fell off rapidly as displacements approached 0.5 km. At lower resolutions, R^2 remained close to 0. For BRTs at 25 m resolution, PR^2 was highest (~0.07) for zero displacement data, and then fell off to near 0.05 at higher displacements. At lower resolutions, PR^2 increased to near 0.1 (at 50 m) and 0.15 (at 75 m), although there was substantial variation around these mean values. Given the amount of variation in PR^2, displacement did not appear to have much influence on the predictive power of BRT models at these resolutions.

Figure 42. Summary of model diagnostics (PR^2) for a series of 50 models run on displaced data in urban areas. Each red point indicates the PR^2 value from a single model run in which data on the proportion of the population under 5 were aggregated at differing resolutions. Black points with error bars indicate the mean and 95% confidence intervals around the mean PR^2 value at each displacement level. Each row indicates the set of results for Generalised Additive Models (GAM), Boosted Regression Trees (BRT) and Linear Models (LM) tested on the same datasets.



6. Discussion

6.1 National-level Geostatistical Mapping

The results presented in this report highlight the potential of geostatistical mapping techniques to produce interpolated surfaces from GPS-located DHS survey variables. Results based on these techniques can help meet the needs of national and international communities for smaller area estimates than those currently provided by The DHS Program.

The histogram and variogram summaries for each indicator (Figure 2, Figure 6, Figure 10 and Figure 14) show that there is a great variation in cluster survey outputs both between variables and between countries, and that each country-variable pair tested is relatively unique in terms of the spatial structure of the data and the accuracy of the mapping undertaken. This has wider implications in terms of extending the approach (and other types of mapping approaches) to other countries and variables, and highlighting that there are no guarantees in terms of consistency in mapping accuracies between countries and variables. Certain variables exhibit much greater ranges of spatial variations and stronger spatial structure, and have spatial structures that are more readily captured by the set of covariates reported here. Conversely, some variables display variation that is apparently random, or at least manifest over spatial scales that are too short to be resolved by the survey data and covariates. Nevertheless, the methods proposed here are designed to be robust to such variations and are explicit in terms of both model validation and mapped uncertainty; this enables users to visualize and explore where mapping accuracy is high and where predictions are uncertain.

Unsurprisingly, the geographic variations in a variable such as access to HIV testing, which has principally non-biophysical drivers, were less able to be captured by the suite of principally environmental covariates available here, with generally lower predictive R^2 values. Despite this, overall absolute errors were relatively low. This showed that the variable was mapped relatively precisely for all three countries with the available covariate suite, and that this performance could be expected to improve as richer geospatial data become available on socio-demographic characteristics that are more closely linked to HIV infection and care.

The covariate selection method produced results that highlighted the key environmental driving factors behind the spatial variation seen in cluster-level outcomes. For instance, elevation was a consistently selected variable in modeling anemia prevalence; accessibility was a key variable in the access to HIV testing and improved sanitation models; and the nightlights data were important in characterizing the urban-rural differences seen in the prevalence of stunting.

A key aspect of the current approach is the challenge of urban area mapping. This is an important issue, given that close to 50% of people in the countries that were mapped reside in urban areas. In all of the national-level predicted maps, urban areas are predicted with relatively uniform values across them. Urban areas, however, typically exhibit substantial heterogeneities in health and development indicators, as evidenced by the large range of indicator values at cluster centroids within urban areas (Figure 2, Figure 6, Figure 10 and Figure 14). However, features of the survey data and covariates used in the analysis make capturing this heterogeneity accurately within urban areas challenging. First, the geospatial covariate layers used here are of insufficient spatial detail to capture significant intra-urban variations. At present, a trade-off still exists between wide-area consistent coverage with satellite-derived and other geospatial layers, and capturing detail at smaller, particularly sub-urban, spatial scales. These trade-offs are beginning to be eroded by newer, finer resolution wide area covariate datasets (see next steps); however, at present they represent a limitation for capturing the relevant heterogeneities exhibited by urban areas. Second, even if such detailed spatial covariates were available, the displacement of cluster centroids by up to 2km in urban

areas can result, for example, in a survey that was conducted in an urban slum being displaced into a high income neighborhood, which would make it impossible to reliably link the survey locations to covariates.

6.2 Effects of Cluster Centroid Displacement

The work presented here has assessed the impact of DHS centroid displacement on the subsequent use of GPS-located survey data in a Bayesian model-based geostatistical framework to generate interpolated surfaces of chosen indicators. As may have been expected, the results were complex, and the extent of the impact of displacement varied between indicators and between individual survey data sets. Nevertheless, a number of overarching conclusions can be drawn. First, the impact of displacement on summary validation statistics (i.e., the overall "performance" of a geostatistical model) was modest, although the effect was more marked where the non-displaced model itself was relatively highly performing. Second, the impact of displacement tended to vary geographically across each interpolated surface; this was apparently driven by a range of factors that include data point density and the heterogeneity of important geographic covariates. The latter point also means that rural areas will, on average, be less affected by impact of displacement, despite being subjected to larger displacement radii because the underlying structural variation in covariate and response data is generally over much larger length scales. These results add to the insights on displacement effects in various other analysis contexts developed previously (Perez-Heydrich et al. 2015; Warren et al. 2015b, 2015a).

In our exploration of the influence of covariate buffering on model fit in urban settings, we found three kinds of results. In many cases (particularly in Ghana), we were not able to fit a meaningful model to the data with the available covariates; in this case, displacement and buffering did not meaningfully affect the near-zero predictive power of the model. This points to the fact that existing sets of high resolution covariates for urban areas that are widely available across all DHS regions are still not comprehensive enough, or capable of capturing relevant heterogeneities to facilitate accurate within-urban mapping. Sometimes, we observed the unexpected result that when models based on the unbuffered data had low (i.e., near 0.02) R^2 values, these values increased with increasing buffer sizes. This was the case for most models with non-negligible R^2 values. The one exception again was anemia in Uganda, where we found relatively high values of R^2 with low buffer sizes that decreased as buffer size increased. This result suggests that factors related to anemia may be more localized, and that displacement has a severe negative effect on the predictive power of models in the urban context. Furthermore, buffering did not appear to improve the outcome in these cases.

These different kinds of responses suggest that the effects of buffering and displacement will vary depending on the situation, and that differing strategies may be needed to preserve the quality of geo-located data in order understand and predict the distribution of health and poverty-related factors within urban areas, while also preserving the anonymity of survey respondents.

Given the many factors that appear to drive the impact of displacement, it is difficult to derive a post-hoc adjustment that could be applied to displaced publically available data in order to explicitly capture this specific additional source of uncertainty. Methods such as regression calibration (Warren et al. 2015a) may be worth exploring in the context of geostatistical models. However, standard application of geostatistical modeling to displaced data would, in part, capture this uncertainty via the degraded spatial autocorrelation structure (although this effect is relatively modest) and the degraded strength of multivariate relationships with covariates.

6.3 The Potential of High Resolution Urban Mapping

For the case study of mapping proportions of the population under 5 years of age (PU5) in part of Kano City, a combination of high resolution covariates and non-linear techniques improved the predictive power

of models over standard linear models such as GLMs. We also showed that household-level data provide much greater predictive power in these contexts than aggregate data. Finally, we showed that the influence of aggregating to larger grid cells and of displacing data by up to 2 km has differing effects on BRT, GAM, and LM models.

First, the finding that aggregation of data leads to lower predictive power in some instances is unsurprising, since the household-level response data represent a much greater sample size. Plots of household-level covariate data show that there appears to be little in the way of (linear) relationships between PU5 and any of the predictor variables used in the analysis. Both BRT and GAM methods are able to exploit nonlinear relationships in the data, particularly where there are skewed or clustered patterns in the response. This feature may be particularly useful when the factors that influence response data are highly localized.

Of the three models tested in this scenario, the BRTs model performed best overall on the basis of retaining the highest R^2 value. This is due in part to the intrinsic ability of BRTs to accommodate non-linear data structures, which had a clear advantage in this case. However, BRTs also contain an element of randomness, which is also likely to contribute to their overall quality, and is related as well to the increased likelihood of selecting a random subset of the covariate data that appear to predict the response data very well. This effect is magnified when larger grid cells are used, since there are fewer response data to predict using the hold-out dataset. This leads to a much greater range of R^2 values and an overall increase in the mean. This effect is present as well with the GAMs and LMs, although it is not as strong, because of the lack of the random splitting element present in the BRTs.

Thus, it appears that BRTs, in combination with sets of high resolution geospatial covariate layers, may offer a good approach to modeling health and poverty-related factors in an urban setting. However, this will require a substantial amount of survey data, since BRT models tend to be data-hungry. Furthermore, our results suggest that while some response data might be robust to displacement, others require high resolution, non-displaced data in order to retain their relationships with covariate data. Displacement or buffering may obscure important trends. Given these results, it may be that additional survey effort may be needed to acquire the sample sizes required for BRTs. This does not address the problem of preserving the anonymity within urban areas. However, where large numbers of people are present, large displacements may not be as necessary, particularly when the output of analysis is aggregated to a large degree.

A final point is that high resolution datasets are increasingly becoming available in urban areas. These include the global human settlement layer (GHSL), Global Urban Footprint (GUF), MODIS and VIIRS-derived datasets at <1km resolution, and a suite of density and distance-based derived layers based on infrastructure data such as those from OpenstreetMap. These data layers may provide greater predictive power in urban settings where spatial change can be sudden.

6.4 Next Steps

In this section, we outline potential refinements to the basic methodological framework established in this report, and then discuss analyses that would demonstrate compelling use or cases for the predictive surfaces that were developed.

6.4.1 Further improvements to the methodology

In this report we have presented a mapping methodology for which multiple future directions exist for improving the accuracy of the mapping approach, extending its scope and utility, and ensuring that mapping is undertaken in an efficient and sustainable fashion. These suggestions are not prescriptive, but are offered as insights into the potential adaption of the methods to differing situations.

Spatial resolution: The exploratory analyses showed that relatively accurate mapping could be undertaken at a resolution of 5 × 5 km. However, in many cases, finer resolutions will be of interest, particularly in urban areas. With relevant geospatial covariate layers continuing to be produced at finer and finer spatial resolutions, there is the potential to explore the benefits of both improved mapping accuracies and map visualizations through modeling at finer resolutions. Denominator population distribution datasets are now being constructed at 100m resolution or finer (e.g., WorldPop, LandScan HD), while covariates on human settlements (e.g., Global Human Settlement Layer (38m resolution), Global Urban Footprint; 17m resolution), infrastructure (e.g., OpenStreetMap) and satellite-derived variables (e.g., sentinel series) are all available at finer spatial resolutions. However, mapping at 1km (or finer) will require explorations of the limits to the improvements in mapping accuracies that can be achieved in the presence of cluster centroid displacements. At finer scales than 1 km, it will be important to investigate other methods such as the boosted regression trees or generalized additive models presented in this report. In addition, the greater volume of data will require development of computational pipelines that allow prediction and manipulation of datasets that exceed the memory capacity of most standard desktop computers.

Additional geospatial covariates: The suite of covariate layers used in the national level analyses presented here represent those currently available and assembled at the University of Oxford for disease mapping work. Continuation of the work should ideally explore both the updating of the existing set of covariates and the testing of new ones, tailored as much as possible to the specific indicator of interest. This may, for example, include updating the accessibility layers to include more recent and detailed road networks, and the detailed settlement layers outlined above. Moreover, if the focus is on the production of the most accurate map for an individual country, a more detailed and comprehensive set of covariate layers exist to enable improved map production, as compared to examples when the goal is consistency between countries. The effect of this tradeoff in loss of accuracy is unclear, and may vary between countries, with effects largest in urban areas where detailed datasets may exist for individual cities (e.g., the Kano City analyses presented here). Refinements of our methods could explore quantitatively the improvements in mapping accuracies obtained through a country-specific focus, rather than the use of globally-consistent covariates. In addition, many socioeconomic and demographic factors that are not captured by the suite of satellite-based covariates used here can be obtained when detailed and contemporary census data are available with sub-city scale information. The potential of incorporating these aggregate level data as covariates where available and assessing their ability to improve mapping accuracies are an area for further exploration.

Mapping change: With many countries now having two or more DHS surveys with geolocated cluster data available, the potential exists to develop methods for quantifying changes geospatially. Challenges exist in undertaking this because of changing cluster locations, changes to survey questions, and changing populations in each country. Nevertheless, monitoring change and progress towards health and development goals sub-nationally could be a valuable application of the geolocated survey data, if robust methods are in place to overcome these challenges and quantify the uncertainty in change estimates. Space-time Bayesian geostatistical methods for undertaking this kind of analysis are being developed by the Malaria Atlas Project, and great potential exists for extending them to other DHS variables.

Integration of other surveys: With multiple aid initiatives unfolding in the developing world, there are a variety of survey types being conducted each year, and sometimes in the same country. The techniques developed here will allow comparison of predictive surfaces from different surveys to detect trends in factors that are not necessarily measured at the same locations. A more challenging, but potentially fruitful, data integration may be between community based households surveys such as DHS and data from routine governmental monitoring systems with data from health facilities, schools, and other civic systems.

Scaling up: This report has shown the potential of the mapping approach for three countries and four variables. If the approach is to be adopted as a standard method for map production and distribution, planning is required to implement scaling up the mapping to multiple other countries and variables. This

will require longer-term planning on methods and organizations for implementation, training, regularity of update, covariate selection and update, variables to be mapped, and data hosting and storage as well as guidance on use of the new map surfaces and the estimates of uncertainty for program planning and decision making. In addition, many more survey variables could potentially be mapped, as suggested in a recent consultative meeting that produced recommendations of DHS survey variables with potential priorities for mapping. Next steps will involve exploration of a larger number and wider range of variables.

6.4.2 Development of use-cases for predictive maps

The availability of maps that predict the prevalence of socio-economic factors at resolutions of 5 km or finer presents a valuable opportunity for sociological research and development applications. However, from an application perspective, we are entering uncharted territory since this granularity of data on health and demography indicators has rarely been available. Described below, we present a number of potential cases for the use of these predictive surfaces in development and health contexts.

Targeting interventions and development: Many public services in developing countries are stretched due to limited resources. Understanding the burden on facilities, such as health facilities and hospitals or where new schools are needed, can be challenging without a detailed picture of the composition of the population. While detailed estimates of population density at fine (100 m) resolution are increasingly available, it is rare for these estimates to contain information on numbers of individuals within specific groups, such as numbers of children with anemia. By using the prevalence surfaces developed here in combination with high-resolution population estimates, it is possible to estimate the total numbers of individuals within certain categories. These predictions could then be used with other GIS measures and survey data, such as the placement of roads or staffing levels at facilities, to determine optimal positioning for new schools or the optimal resources needed in a new health facility. Similarly, resources such as bed nets or vaccines can be targeted to areas of the greatest need.

Improving burden estimates: Currently, national level or coarse administrative unit level mapping masks heterogeneities and misses possible hotspots and inequalities. There is potential for our methods to better identify these and to work with population maps to more accurately quantify and map burdens of disease and other health related issues. For example, this approach has been used in the context of estimating the burden of malaria (Hay and Snow 2006), as well as the distribution of children and women of childbearing age (Tatem et al. 2013).

Understanding deviations from the norm: With widespread health and social problems, it can be difficult to determine which areas are better off than others. In addition, it can be difficult to discern whether new observations at a location are anomalous or within expected bounds. The uncertainty measures provided in our predictive surfaces provide a means by which one can evaluate whether observed differences between locations are meaningful in the context of the broader scale. For example, new observations of anemia prevalence could be compared against the background surface to determine if they represent a higher or lower prevalence than expected for that area.

Multi-temporal analyses: The initial analyses presented here focused on exploring the potential for mapping at individual time points that matched the survey dates. With The DHS Program undertaking repeated surveys for many countries, however, there exists potential to make use of multiple surveys across time to improve the accuracy of output maps, and to produce separate maps for different time points and quantify change. Space-time geostatistical methods being constructed at the University of Oxford to quantify and attribute changing malaria prevalences could be adapted here. Given that current Millennium Development Goals and Sustainability Goals are based on measurable outcomes, this ability to measure change over time will be increasingly important.

7. Conclusion

This study has presented and tested a flexible and robust geostatistical framework for generating and validating interpolated surfaces by using georeferenced DHS survey data across a variety of indicator and country settings. The ultimate precision of interpolated surfaces that can be achieved varies between settings, and is driven by the spatial structure of each indicator and its relationship to available covariate data. The random displacement of DHS cluster geopositioning information, used to protect respondent anonymity, tends to reduce the precision of predicted maps, although the impact varies between settings and is generally modest. Over shorter distances, the greater degree of geographical heterogeneity associated with urban areas means that these areas present particular challenges to accurate geospatial mapping. They are also more sensitive to the impact of cluster displacement. High resolution covariates and novel statistical approaches show potential to improve mapping in these areas.

There are a number of prominent examples of the use of DHS survey data for generating interpolated surfaces. This study has demonstrated that, provided that appropriate modeling and validation are carried out, such data have a broad utility for creating maps of a wide range of indicators that support improved geographically stratified decision making.

References

Antosiewicz, H.A. 1964. "Bessel Functions of Integer Order." In *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, edited by M. Abramowitz and I. A. Stegun, 435-478. New York, U.S.A.: Dover Publications Inc.

Blangiardo, M., M. Cameletti, G. Baio, and H. Rue. 2013. "Spatial and Spatio-temporal Models with R-INLA." *Spatial and Spatio-temporal Epidemiology* 7:39-55.

Burgert, C.R. 2014. "Spatial Interpolation with Demographic and Health Survey Data: Key Considerations." *DHS Spatial Analysis Reports No. 9*. Rockville, Maryland, USA: ICF International. Available at http://dhsprogram.com/pubs/pdf/SAR9/SAR9.pdf.

Burgert, C.R., J. Colston, T. Roy, and B. Zachary. 2013. "Geographic Displacement Procedure and Georeferenced Data Release Policy for the Demographic and Health Surveys" *DHS Spatial Analysis Reports No. 7* Calverton, Maryland, USA: ICF International.

Davis, G.M. 1964. "Gamma Function and Related Functions." In *Handbook of Mathematical Functions: with Formulas, Graphs, and Mathematical Tables*, edited by M. Abramowitz and I. A. Stegun, 253-293. New York, U.S.A.: Dover Publications Inc.

Diggle, P.J., and P.J. Ribeiro. 2007. *Model-based Geostatistics*. Edited by P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin and S. Zeger, *Springer Series in Statistics*. New York: Springer.

Diggle, P.J., J.A. Tawn, and R.A. Moyeed. 1998. "Model-based Geostatistics." *Journal of the Royal Statistical Society Series C-Applied Statistics* 47:299-326.

Elith, J., J.R. Leathwick, and T. Hastie. 2008. "A Working Guide to boosted Regression Trees." *Journal of Animal Ecology* 77 (4):802-813.

Fong, Y., H. Rue, and J. Wakefield. 2009. "Bayesian Inference for Generalized Linear Mixed Models." *Biostatistics*:kxp053.

Gething, P.W., I.R.F. Elyazar, C.M. Moyes, D.L. Smith, K.E. Battle, C.A. Guerra, A.P. Patil, A.J. Tatem, R.E. Howes, M.F. Myers, D.B. George, P. Horby, H.F.L. Wertheim, R.N. Price, I. Müeller, J.K. Baird, and S.I. Hay. 2012. "A Long Neglected World Malaria Map: *Plasmodium Vivax* Endemicity in 2010." *PLoS Neglected Tropical Diseases* 6:e1814.

Gething, P.W., A. Patil, D. Smith, C. Guerra, I. Elyazar, G. Johnston, A. Tatem, and S. Hay. 2011. "A New World Malaria Map: *Plasmodium Falciparum* Endemicity in 2010." *Malaria Journal* 10 (1):378.

Hay, S.I., and R.W. Snow. 2006. "The Malaria Atlas Project: Developing Global Maps of Malaria Risk." *PLoS Med* 3 (12):e473.

ICF International. 2008-2011. *Demographic and Health Surveys* (various). Rockville, Maryland: ICF International.

Perez-Heydrich, C., J. Warren, C. Burgert, and M. Emch. 2015. "Influence of Demographic and Health Survey Point Displacements on Raster-based Analyses." *Spatial Demography*:1-19.

Rue, H., S. Martino, and N. Chopin. 2009. "Approximate Bayesian Inference for Latent Gaussian Models by Using Integrated Nested Laplace Approximations." *Journal of the Royal Statistical Society: Series b (Statistical Methodology)* 71 (2):319-392.

Tatem, A.J., A.J. Garcia, R.W. Snow, A.M. Noor, A.E. Gaughan, M. Gilbert, and C. Linard. 2013. "Millennium Development Health Metrics: Where do Africa's Children and Women of Childbearing Age Live?" *Popul Health Metrics* 11 (1):11.

Tatem, A.J., P.W. Gething, C. Pezzulo, S. Bhatt, and D. Weiss. 2014. *Final Report: Development of High-resolution Gridded Poverty Surfaces, Bill and Melinda Gates Foundation Contract #21989*.

Warren, J., C. Perez-Heydrich, C. Burgert, and M. Emch. 2015a. "Influence of Demographic and Health Survey Point Displacements on Distance-Based Analyses." *Spatial Demography*:1-19.

Warren, J., C. Perez-Heydrich, C. Burgert, and M. Emch. 2015b. "Influence of Demographic and Health Survey Point Displacements on Point-in-Polygon Analyses." *Spatial Demography*:1-17.