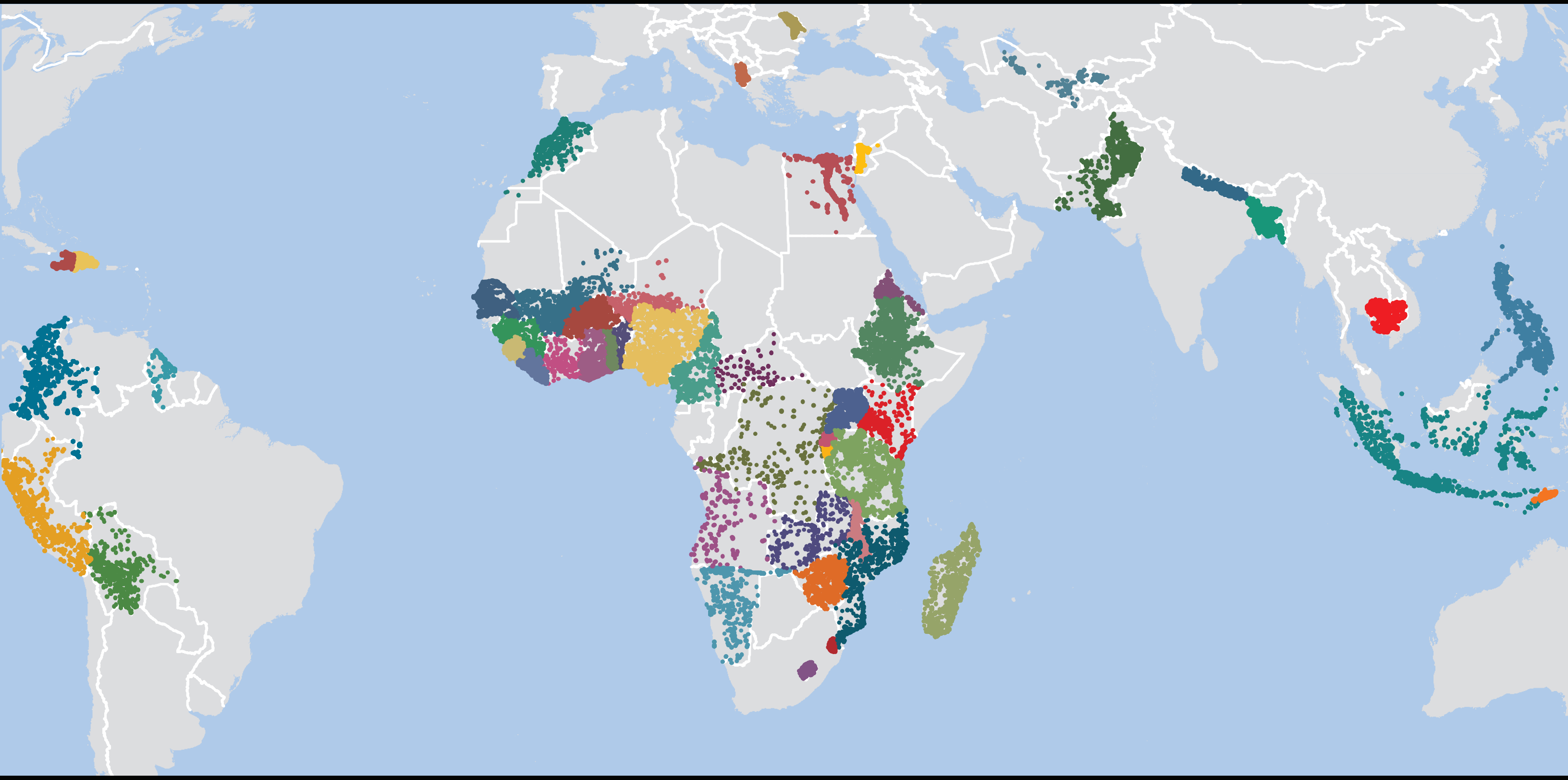
USAID
FROM THE AMERICAN PEOPLE

THE DHS PROGRAM MODELED MAP SURFACES: UNDERSTANDING THE UTILITY OF SPATIAL INTERPOLATION FOR GENERATING INDICATORS AT SUBNATIONAL ADMINISTRATIVE LEVELS

DHS SPATIAL ANALYSIS REPORTS 15



August 2017

This publication was produced for review by the United States Agency for International Development (USAID). The report was prepared by Peter W. Gething and Clara R. Burgert-Brucker.

DHS Spatial Analysis Reports No. 15

The DHS Program Modeled Map Surfaces: Understanding the Utility of Spatial Interpolation for Generating Indicators at Subnational Administrative Levels

Peter W. Gething

Clara R. Burgert-Brucker

ICF
Rockville, Maryland, USA

August 2017

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

Acknowledgment:
The authors would like to acknowledge Trinadh Dontamsetti and Tom Pullum for their assistance in finalizing this report, and Emanuele Giorgi for the review of this report.

Editor: Bryant Robey
Document Production: Teresa Duberry

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, 530 Gaither Road, Suite 500, Rockville, MD 20850, USA. phone: 301-407-6500; fax: 301-407-6501; email: reports@dhsprogram.com; Internet: www.dhsprogram.com.

Recommended citation:

Gething, Peter W. and Clara R. Burgert-Brucker. 2017. *The DHS Program Modeled Map Surfaces: Understanding the Utility of Spatial Interpolation for Generating Indicators at Subnational Administrative Levels*. DHS Spatial Analysis Reports No. 15. Rockville, Maryland, USA: ICF.

Contents

Tables

Figures

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 this series 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

As interest in describing subnational variation in demographic and health indicators grows, The DHS Program has commenced a plan of work to enable the routine creation and dissemination of spatially modeled indicator surfaces for a set of key indicators. This will be done based on model-based geostatistical (MBG) techniques. To further understand the performance of these models and the factors influencing that performance, this study has investigated (1) what common factors drive MBG modeled surface accuracy and (2) how accurately MBG models can predict aggregated estimates at a finer spatial scale than the first subnational administrative unit level (SNU1).

To explore factors influencing model performance, 15 DHS indicators were modeled across 16 countries. Certain indicators performed consistently well across all countries, and others consistently less well. The amount of cluster-level variation in each indicator and the extent to which that variation was spatially autocorrelated were the two most important factors in determining the accuracy of modeled surfaces. It was possible to predict how accurately a given indicator could be mapped based on simple attributes of the raw input data.

To explore the ability of MBG models to provide accurate indicator estimates below the SNU1 level, experiments were conducted using the 2014 Kenya DHS, which is unique in having been sampled with approximately four times greater density of clusters than a standard DHS. Analyses based on artificially thinned versions of this data demonstrated that MBG surfaces and aggregated estimates performed progressively better as survey sample size was increased. We found that the use of a geostatistical model to estimate aggregated indicator estimates tends to yield more precise estimates than the default approach of directly calculating weighted means of survey data. On average across the indicators and performance metrics, the use of a geostatistical model based on a standard DHS survey to estimate indicators at "standard" SNU2 level (i.e., at a level of geographical aggregation equivalent to SNU2 in most countries) yields results of accuracy equivalent to a survey three times larger in the absence of a geostatistical model.

The integration of geospatial methods in the survey design and subsequent data analysis stages should be considered for future DHS surveys, especially if a desired outcome is to provide precise estimates below the SNU1 level. This integration will provide more precise estimates below the SNU1 level and do so with fewer requirements for large sample sizes.

Executive Summary

Background and Objectives

Improved understanding of geographic variation in demographic and health indicators within countries is increasingly recognized as central to meeting development goals but, when assessed only at national or highly aggregated subnational levels, important geographical inequities are often concealed. In an international policy environment with limited funding for health and development, the ability to target limited resources to populations of highest need becomes crucial. To meet the growing need for indicator estimates at much finer levels of geographic disaggregation, DHS has commenced a program of work to enable the routine creation and dissemination of spatially modeled indicator surfaces for a set of key indicators, based on model-based geostatistical (MBG) techniques. This increasing interest in modeled surfaces has motivated further questions around the factors that determine their accuracy in different settings, and the extent to which the techniques can form the basis for aggregated estimates at small subnational decision making units. Accordingly, the current report has two main objectives:

Objective 1: to investigate whether common factors can be identified that drive the accuracy with which a modeled indicator surface can be generated using MBG approaches.

Objective 2: to explore how accurate the surfaces are when used to generate aggregate indicator estimates below (i.e., at finer spatial scale) the first-level subnational units (SNU1) for which standard DHS surveys are ostensibly designed.

Understanding Determinants of Accuracy

Investigating the potential drivers of geospatial model accuracy was conducted in four parts. First, we used model-based geostatistics to generate modeled spatial surfaces for a large set of indictors and country surveys. Second, out-of-sample validation was performed for each of these country-indicator surfaces and accuracy metrics were calculated. Third, a set of statistics was computed for each country-indicator that described a broad range of attributes associated with each survey and indicator, and we examined the correspondence between those attributes and the accuracy of the subsequent modeled spatial surface. Fourth, an exercise was conducted to assess the extent to which, given only knowledge of those descriptive data attributes, the potential accuracy of a modeled spatial surface could be predicted.

Comparison of model performance across 16 countries and 15 indicators showed considerable variation. In general, the performance for a given indicator was somewhat consistent across countries: indicators such as ITN access, improved sanitation and female literacy, for example, were predicted with consistently good accuracy across all countries while others, such as child vaccination, were consistently less accurate. Indicator type had more influence on performance than did country/survey setting, although some surveys did tend to perform better than others across all indicators.

Analysis of the factors driving model performance revealed that important factors included the amount of variation in the indicator observations between survey clusters, and the extent to which that variation was spatially autocorrelated. A meta-model that aimed to predict the likely performance of a geospatial model based only on attributes of the input data had good predictive power, able to predict future performance statistics with an R2 of around 0.75. This means that, when deciding which indicators to potentially subject to a geostatistical mapping exercise, it should be possible to predict reasonably well whether the resulting maps will or will not meet a particular level of accuracy.

Understanding Utility of Spatial Modeling for Generating Indicators at Second-Level Subnational Administrative Levels

Following a national reorganization of administrative unit boundaries, the 2014 Kenya DHS was designed with a uniquely dense sampling frame, approximately four times larger than a traditional DHS survey. This provides a unique opportunity to test the ability of geostatistical modeled surfaces to generate sufficiently accurate indicator estimates at what, in nearly every country other than Kenya, would be considered the SNU2 level. The approach proceeded in four main steps. First, the geospatial model was used to generate County-level indicator estimates based on the full 2014 Kenya DHS data set, and these were compared with the directly calculated (i.e. weighted mean) estimates presented in the 2014 Kenya DHS report. For later reference, pixel-level modeled surfaces were also produced for each indicator using the full data set. Second, the full 2014 Kenya DHS data set was then artificially thinned by progressive intervals, with the most thinned version replicating the survey design of a standard DHS—based on the 2008 Kenya DHS. Third, the thinned data sets were then used to generate both pixel-level and County-level geospatial estimates, and the effects of thinning on performance were measured. Fourth, the relative accuracy of County-level estimates derived from thinned data sets was compared between the geospatial model and directly calculated (weighted mean) estimates.

We found that the use of a geostatistical model to estimate aggregated indicator estimates tends to yield more precise estimates than the default approach of directly calculating weighted means of survey data. On average across the indicators and performance metrics, the use of a geostatistical model based on a standard DHS survey to estimate indicators at "standard" SNU2 level (i.e. at a level of geographical aggregation equivalent to SNU2 in most countries) yields results of equivalent accuracy to a survey three times larger in the absence of a geostatistical model.

Implications and Recommendations

An immediate and general recommendation arising from this work is that the integration of geospatial methods in the survey design and subsequent data analysis stages should be considered for future DHS surveys, especially if a desired outcome is to provide precise estimates below the SNU1 level. This integration will provide more precise estimates below the SNU1 level and do so with fewer requirements for large sample sizes.

To fully operationalize this recommendation, the analyses presented here would require a number of extensions. First, repeating the work implemented here for the 2014 Kenya DHS to at least one other survey with similarly dense sampling would provide verification of the generalizability of the findings. Second, a generalized framework could be envisioned that would allow prospective survey design to be carried out to achieve pre-specified levels of precision using the geospatial model, taking into account the exact nature of SNU2 units in a given country. Such a framework would require the current analysis to be repeated for a synthetic set of administrative units at progressively smaller levels of aggregation, to provide reference results against which actual units in other countries could be compared.

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. When assessed only at national or highly aggregated subnational levels, development and health statistics can often conceal important geographical inequities. In an international policy environment with limited funding for health and development, the ability to target limited resources to populations of highest need becomes crucial.

The Demographic and Health Survey (DHS) Program has been a leader in collecting and providing cluster-randomized survey data on core development indicators, traditionally described in survey reports with indicator statistics disaggregated at first-order subnational regions (for example at province or state level) and urban-rural strata. In the context of DHS surveys, a cluster is usually defined based on a stratified two-stage cluster design with a first stage selecting Enumeration Areas (EAs), generally drawn from national census files, and a second stage sampling households within each EA from a household list. The availability in most recent surveys of GPS coordinates for DHS clusters—as well as for Malaria Indicator Surveys (MIS) and AIDS Indicator Surveys (AIS) clusters—provides highly resolved locational information that can be linked with survey outputs for quantifying demographic and health status heterogeneities and inequities.

To meet the growing need for indicator estimates at much finer levels of geographic disaggregation, DHS has commenced a program of work to enable the routine creation and dissemination of spatially modeled indicator surfaces for a set of key indicators to accompany current and future population-based DHS surveys and for a selection of earlier surveys. The maps are publicly available for download on the DHS Spatial Data Repository (spatialdata.dhsprogram.com). This work has sought to deepen understanding of the challenges and best practice for the generation of modeled spatial surfaces based on DHS survey data, and to provide practical guidance to potential users.

This process began in June 2013 with a meeting convened by DHS to bring together key stakeholders to discuss the use of geographic data from DHS population-based surveys for spatial interpolation. DHS Spatial Analysis Report 9 (SAR 9) (Burgert 2014) summarizes key discussions and recommendations from that meeting including indicator selection, methods, and limitations. Following the June 2013 meeting, DHS began exploring the potential use of Bayesian model-based geostatistics (MBG) for the production of interpolated modeled surfaces from the DHS population-based survey GPS-located cluster data, testing MBG methods on four indicators in three surveys. DHS Spatial Analysis Report 11 (SAR 11) (Gething et al. 2015) summarizes the detailed results of this proof of concept activity, including the assessment of method validity, covariates, and uncertainty. The impact on MBG spatially modeled surfaces of the geo-masking of DHS cluster coordinates was also investigated. Following that proof-of-concept, DHS moved to routine production and dissemination of spatially modeled surfaces, and a detailed accompanying guidance document for end users is published as DHS Spatial Analysis Report 14 (SAR 14) (Burgert et al. 2016).

With this growing suite of completed modeled surfaces, each with accompanying validation statistics assessing model performance, it is notable that the precision with which surfaces can be generated (when assessed against cluster-level indicator data in out-of-sample tests) varies considerably both between indicators and between country surveys. The current analysis was conceived to further understand the drivers and implications of this varying accuracy, with two distinct objectives.

1.1 Objective 1: Understanding factors determining accuracy of spatially interpolated surfaces

Objective 1 is to investigate whether common factors can be identified that drive the accuracy with which a modeled indicator surface can be generated using MBG approaches. This leverages the large suite of indicators and country surveys that have now been subject to a standardized MBG approach to generate modeled surfaces, with consistent out-of-sample validation procedures to measure performance and accuracy. Understanding the drivers of accuracy is useful for a number of reasons. First, it can define useful rules-of-thumb as to indicators or country settings where DHS survey data are likely to be more or less amenable to generation modeled surfaces with high accuracy. Second, it potentially informs future survey design to optimize the utility of the resulting data for modeled surface creation.

1.2 Objective 2: Understanding utility of spatial modeling for generating indicators at second-level subnational administrative levels

Objective 2 is to explore how accurate the surfaces are when used to generate aggregate indicator estimates below (i.e. at finer spatial scale) the first-level subnational units (SNU1) for which standard DHS surveys are ostensibly designed. In a standard DHS design it is not logically possible to evaluate the accuracy of aggregate estimates, whether at SNU1 or second-level subnational units (SNU2), because there is no other set of gold-standard values against which to compare. In Kenya, recent changes to the national system of subnational administrative units have meant that the 2014 Kenya DHS survey was uniquely densely sampled—with approximately a four times greater density of survey clusters than a standard DHS. This presented a unique opportunity to provide a gold-standard set of subnational indicator values against which geospatial estimates can be tested.

1.3 Report structure

The remainder of this report describes the analyses conducted to address these two objectives, presents the results, and discusses the implications. Section 2 describes in more detail the DHS surveys and indicators included in this analysis. Section 3 then presents the analysis, results, and discussion for Objective 1, and Section 4 presents these for Objective 2.

2 DHS Surveys and Indicators

Fifteen indicators were analyzed in this project, they came from 16 national DHS surveys for which modeled surfaces were created by the DHS in 2016. The resulting potential set of 240 survey-indicator pairs (i.e., 15×16) was reduced to 233, because certain indicators—for example, the prevalence of anemia in children—were not obtained in all surveys. Table 1 lists the DHS surveys included in this study (referred to in this report as country-surveys), and Table 2 lists the indicators addressed, along with their formal definitions.

Table 1. DHS surveys included in this study, with abbreviated Dataset ID codes used throughout this document

Country	Year	Dataset ID	Full DHS survey name
Bangladesh	2014	BD2014DHS	Bangladesh 2014 Demographic and Health Survey
Cambodia	2014	KH2014DHS	Cambodia 2014 Demographic and Health Survey
Democratic Republic of the Congo	2013-14	CD2013DHS	Democratic Republic of the Congo 2013-14 Demographic and Health Survey
Dominican Republic	2013	DR2013DHS	Dominican Republic 2013 Demographic and Health Survey
Egypt	2014	EG2014DHS	Egypt 2014 Demographic and Health Survey
Ghana	2014	GH2014DHS	Ghana 2014 Demographic and Health Survey
Liberia	2013	LB2013DHS	Liberia 2013 Demographic and Health Survey
Mali	2012-13	ML2012DHS	Mali 2012-13 Demographic and Health Survey
Nigeria	2013	NG2013DHS	Nigeria 2013 Demographic and Health Survey
Namibia	2013	NM2013DHS	Namibia 2013 Demographic and Health Survey
Rwanda	2015	RW2015DHS	Rwanda 2015 Demographic and Health Survey
Sierra Leone	2013	SL2013DHS	Sierra Leone 2013 Demographic and Health Survey
Togo	2013-14	TG2013DHS	Togo 2013-14 Demographic and Health Survey
Zambia	2013-14	ZM2013DHS	Zambia 2013-14 Demographic and Health Survey

Table 2. DHS indicators addressed in the study, their definitions, and abbreviated ID codes used throughout this document

Indicator	Definition	Indicator ID
Married women currently using any modern method of contraception	Percentage of currently married or in union women currently using any modern method of contraception	FP_CUSM_W_MOD
Demand for family planning satisfied by modern methods	The number of currently married women using modern methods of family planning divided by the number of currently married women with demand for family planning (either with unmet need or currently using any family planning)	FP_NADM_W_PDM
Unmet need for family planning	Percentage of currently married or in union women with an unmet need for family planning	FP_NADM_W_UNT
Antenatal visits for pregnancy: 4+ visits	Percentage of women who had a live birth in the 5 years preceding the survey who had 4+ antenatal care visits	RH_ANCN_W_N4P
Place of delivery: Health facility	Percentage of live births in the 5 years preceding the survey delivered at a health facility	RH_DELP_C_DHF
Women who are literate	Percentage of women age 15-49 who are literate	ED_LITR_W_LIT
DPT3 vaccination received	Percentage of children age 12-23 months who had received a third dose of DPT	CH_VACC_C_DP3
Measles vaccination received	Percentage of children age 12-23 months who had received measles vaccination	CH_VACC_C_MSL
Men who are literate	Percentage of men age 15-49 who are literate	ED_LITR_M_LIT
Tobacco use among men	Percentage of men age 15-49 who use tobacco	100 -AH_TOBC_M_NON
Population living in households using an improved water source	Percentage of the de jure population living in households whose main source of drinking water is an improved source	WS_SRCE_P_IMP
Population living in households using no toilet facility	Percentage of the de jure population living in households whose main type of toilet facility is no facility (practicing open defecation)	WS_TLET_P_NFC
Persons with access to an ITN	Percentage of the de facto household population who could sleep under an ITN if each ITN in the household were used by up to two people	ML_ITNA_P_ACC
Women age 15-49 with any anemia	Percentage of women classified as having any anemia (<12.0 g/dl for non-pregnant women and <11.0 g/dl for pregnant women)	AN_ANEM_W_ANY
Children stunted	Percentage of children under age 5 stunted (below -2 SD of height-for-age according to the WHO standard)	CN_NUTS_C_HA2

3 Understanding Determinants of Accuracy

This section describes the work undertaken to address Objective 1, as defined in section 1.1—to investigate whether common factors can be identified that drive the accuracy with which a modeled indicator surface can be generated using MBG approaches. Methods are described first, followed by results and discussion.

3.1 Methods

Investigating the potential drivers of geospatial model accuracy was conducted in four parts. First, we used model-based geostatistics to generate modeled spatial surfaces for a large set of indictors and country surveys. Second, out-of-sample validation was performed for each of these country-indicator surfaces and accuracy metrics were calculated. Third, a set of statistics was computed for each country-indicator pair that described a broad range of attributes associated with each survey and indicator, and we examined the correspondence between those attributes and the accuracy of the subsequent modeled spatial surface. Fourth, an exercise was conducted to assess the extent to which, given only knowledge of those descriptive data attributes, the potential accuracy of a modeled spatial surface could be predicted. This section now describes the methodology for each of these components in more detail.

3.1.1 Model-based geostatistical modeling of DHS indicators

The Bayesian model-based geostatistical (MBG) framework for generating standardized modeled surfaces for DHS indicators has been described in detail in SAR 11 (Gething et al. 2015) and SAR 14 (Burgert et al. 2016). A summary description of the approach is included in this report as Annex 2. The study used this approach to generate a 5x5km pixel resolution raster surface for each of the 233 country-indicator pairs listed in section 2.

3.1.2 Comparing validation performance statistics across countries and indicators

3.1.2.1 Out-of-sample validation procedure

For each of the 233 country-indicator model outputs, a validation procedure was implemented and a set of performance statistics was calculated. This proceeded using an out-of-sample validation consisting of a four-fold hold-out procedure, whereby 25% of the data points were randomly withdrawn from the dataset; the model was run in full using the remaining 75% of data, and the predicted values at the locations of the hold-out data were compared to their observed values. This procedure was repeated four times without replacement such that every data point was held out once across the four validation runs. Standard validation statistics were then computed as measures of the predictive accuracy of the modeled estimates:

- **Correlation (COR)**: degree of linear association between observed and predicted values.

- **Mean absolute error (MAE)**: quantifies model precision—i.e. the average magnitude of difference between observed and predicted values. This is computed in the same units as the variable being predicted (so, if the indicator is a rate expressed on a scale from 0-100%, then the MAE will also be a value from 0-100%).

- **Mean square error (MSE)**: indicates how accurate the model is, encapsulating bias and error, with values close to 0 providing an indication that the model is more accurate, and values close to 1 indicating the model is less accurate.

The current study has focused on accuracy of the point estimate as the most important aspect of model performance. In some contexts, other performance aspects are also important, but these are not considered here. One such measure is the extent to which the model-generated metrics of uncertainty accompanying these point estimates capture their true uncertainty. This can be evaluated, for example, by calculating the fraction of held-out observations that fall within a given uncertainty interval.

3.1.2.2 Exploring variation in accuracy

Once all validation statistics had been computed and assembled, a number of visualizations were generated to allow exploratory assessment of patterns of variation in modeled indicator accuracy. First, a set of boxplots was generated that summarized the variation in each of the three performance metrics (COR, MAE, MSE) by indicator and by country-survey. Second, for each performance metric, color-coded matrices were generated that tabulated country-surveys against indicators such that each cell contained the performance statistic for a given country-indicator combination, with cells colored proportional to value. These matrices allow visualization of consistent patterns along rows (i.e., where a given indicator has consistently high/low performance across most surveys or, conversely, where a given country-survey has consistently high/low performance across most indicators).

3.1.3 Analysis of correspondence between model performance and explanatory factors

3.1.3.1 Defining and measuring data attributes

For each country-indicator pair, eight characteristics of the cluster-level indicator data were defined and measured, and subsequently explored for their effect on modeled surface accuracy. These were as follows.

Point density: the national-level density of survey clusters expressed as points per km^2.

N per cluster: the national-level mean number of individuals per cluster included in the denominator for a given indicator.

Cluster-level mean: the mean value of the indicator across all survey clusters.

Cluster-level variance: the variance of the indicator across all survey clusters.

Spatial variance: the magnitude of variance of the indicator across all survey clusters that is spatially autocorrelated. This was computed using the partial sill parameter from a fitted variogram model.

Range of spatial correlation: the geographical distance over which the indicator is positively autocorrelated (i.e., the maximum distance over which two observed indicator values are likely to display smaller variance than two observations at an arbitrarily large separation distance). This was estimated using the range parameter from a fitted variogram model.

Non-spatial variance: the magnitude of variance of the indicator across all survey clusters that is not spatially autocorrelated. This was computed using the nugget parameter from a fitted variogram model.

Spatial variance fraction: the fraction of cluster-level variance that is spatially autocorrelated. This was calculated simply as the spatial variance divided by spatial plus non-spatial variance.

3.1.3.2 Evaluating correspondence between data attributes and modeled surface accuracy

Bivariate regression models were constructed to enumerate the association between each data attribute and each performance metric. Each regression was therefore based on 233 paired observations of a given data attribute and a given performance metric, one from each country-indicator combination. Visual inspection of the scatterplots suggested that relationships were non-linear in many cases, and so a simple polynomial regression was used, with the resulting R^2 value recorded.

3.1.4 Meta-modeling to predict performance a priori

In addition to investigating separately the effects of each of the different data attributes on modeled surface accuracy, we conducted an analysis to consider their combined effects. This was intended to explore the possibility of being able to predict, before undertaking any geospatial analysis, the likely accuracy of any modeled surface, given only knowledge of key attributes of the cluster-level data. To evaluate this, a multivariate regression was conducted for each of the three performance metrics with all eight data attributes included as dependent variables. The fitted coefficients and resulting R^2 values were then extracted and tabulated.

3.2 Results

3.2.1 Validation performance statistics across countries and indicators

Table 3 provides the results of the out-of-sample validation exercise presenting pixel-level performance statistics (COR, MAE, MSE) for every country-indicator, while Figure 1 shows the boxplots summarizing the variation in these statistics across country-surveys and across indicators. Comparison of the left column of boxplots (variation between countries, plotted for each indicator) and right (variation between indicators, plotted for each country) shows that the magnitude of variation is broadly similar. In other words, the accuracy of modeled surfaces varies by a similar amount across indicators in any given country as it does across countries for any given indicator. Some indicators performed consistently well in all country-surveys, with ITN access, improved sanitation and water source, and female literacy tending to show above average performance (higher correlation, lower MAE and MSE) across all country-surveys. Conversely, the child vaccination indicators were consistently below average across all metrics and all countries. When comparing country-surveys, the picture was generally more mixed: most countries displayed a fairly broad range of performance across the different indicators.

Figure 2, Figure 3, and Figure 4 show the color-coded matrices detailing, respectively, COR, MAE, and MSE for each country-indicator combination. Patterns of consistent blue or red shades along rows highlight country-surveys that performed consistently well (blue) or poorly (red) relative to others. For COR, the Nigeria 2013 DHS performed consistently well, for example, while the Dominican Republic 2013 DHS fared consistently poorly. For MAE and MSE there were less consistent patterns of variation between countries. When assessing by column, patterns of consistent blue or red shades highlight indictors that performed consistently well (blue) or poorly (red) relative to others. Corroborating the patterns in Figure 1, ITN access demonstrated consistently high COR and low MAE and MSE, and the child vaccination indicators the opposite.

Table 3 (part 1). Out-of-sample performance statistics (COR=correlation; MAE=Mean Absolute Error; MSE = Mean Square Error) for geostatistical models predicting 15 DHS indicators (columns) for 16 DHS surveys (rows). See Table 1 and Table 2, respectively, for full details of survey and indicator ID codes.

		100 -AH_TOBC_M_NON	AN_ANEM_W_ANY	CH_VACC_C_DP3	CH_VACC_C_MSL	CN_NUTS_C_HA2	ED_LITR_M_LIT	ED_LITR_W_LIT	FP_CUSM_W_MOD	FP_NADM_W_PDM	FP_NADM_W_UNT	ML_ITNA_P_ACC	RH_ANCN_W_N4P	RH_DELP_C_DHF	WS_SRCE_P_IMP	WS_TLET_P_NFC
CD2013DHS	COR	0.770	0.620	0.830	0.670	0.740	0.670	0.870	0.710	0.620	0.590	0.880	0.730	0.880	0.770	0.820
	MAE	0.100	0.120	0.150	0.180	0.090	0.090	0.100	0.040	0.090	0.090	0.070	0.130	0.090	0.170	0.070
	MSE	0.020	0.020	0.040	0.050	0.010	0.010	0.020	0.000	0.010	0.010	0.010	0.020	0.020	0.080	0.010
BD2014DHS	COR			0.480	0.490	0.590	0.800	0.660	0.680	0.630	0.630		0.660	0.720	0.800	0.680
	MAE			0.180	0.200	0.120	0.110	0.100	0.080	0.090	0.050		0.160	0.160	0.030	0.020
	MSE			0.040	0.050	0.020	0.020	0.020	0.010	0.010	0.000		0.040	0.040	0.010	0.000
DR2013DHS	COR	0.430		0.510	0.530	0.450	0.670	0.670	0.430	0.370	0.380		0.390	0.410	0.670	0.660
	MAE	0.070		0.280	0.300	0.100	0.070	0.060	0.120	0.110	0.090		0.110	0.080	0.060	0.030
	MSE	0.010		0.100	0.120	0.010	0.010	0.010	0.020	0.020	0.010		0.020	0.010	0.010	0.000
EG2014DHS	COR		0.450	0.250	0.380	0.540	0.800	0.670	0.480	0.440	0.310		0.460	0.610	0.820	0.070
	MAE		0.210	0.190	0.320	0.160	0.110	0.120	0.150	0.150	0.100		0.140	0.120	0.030	0.010
	MSE		0.070	0.040	0.130	0.050	0.020	0.020	0.040	0.040	0.020		0.030	0.020	0.000	0.000
GH2014DHS	COR	0.630	0.560	0.520	0.500	0.540	0.800	0.850	0.580	0.550	0.520	0.720	0.700	0.850	0.780	0.890
	MAE	0.070	0.120	0.190	0.180	0.130	0.110	0.100	0.100	0.160	0.100	0.080	0.090	0.100	0.080	0.090
	MSE	0.010	0.020	0.050	0.040	0.030	0.020	0.020	0.020	0.040	0.020	0.010	0.010	0.020	0.030	0.020
KE2014DHS	COR	0.610		0.500	0.530	0.530	0.800	0.940	0.830	0.710	0.520	0.880	0.620	0.830	0.790	0.930
	MAE	0.130		0.200	0.200	0.110	0.110	0.060	0.110	0.200	0.070	0.090	0.140	0.140	0.150	0.050
	MSE	0.030		0.050	0.050	0.020	0.020	0.010	0.020	0.060	0.010	0.010	0.030	0.030	0.050	0.010
KH2014DHS	COR	0.570	0.630	0.590	0.590	0.540	0.550	0.770	0.570	0.590	0.560		0.730	0.720		0.770
	MAE	0.140	0.090	0.220	0.240	0.130	0.110	0.090	0.100	0.120	0.060		0.120	0.110		0.140
	MSE	0.030	0.010	0.060	0.080	0.030	0.020	0.010	0.020	0.020	0.010		0.020	0.020		0.040
LB2013DHS	COR	0.730		0.720	0.620	0.570	0.750	0.870	0.660	0.640	0.560	0.810	0.710	0.800	0.710	0.760
	MAE	0.070		0.180	0.190	0.120	0.110	0.090	0.080	0.130	0.090	0.080	0.110	0.120	0.180	0.170
	MSE	0.010		0.050	0.050	0.020	0.020	0.010	0.010	0.030	0.010	0.010	0.020	0.020	0.070	0.060

Table 3 (part 2). See part 1 for caption.

		100 -AH_TOBC_M_NON	AN_ANEM_W_ANY	CH_VACC_C_DP3	CH_VACC_C_MSL	CN_NUTS_C_HA2	ED_LITR_M_LIT	ED_LITR_W_LIT	FP_CUSM_W_MOD	FP_NADM_W_PDM	FP_NADM_W_UNT	ML_ITNA_P_ACC	RH_ANCN_W_N4P	RH_DELP_C_DHF	WS_SRCE_P_IMP	WS_TLET_P_NFC
LS2014DHS	COR	0.550	0.550	0.570	0.540	0.520	0.670	0.550	0.620	0.490	0.470		0.520	0.630	0.480	0.830
	MAE	0.150	0.110	0.240	0.220	0.180	0.120	0.040	0.130	0.140	0.130		0.130	0.140	0.160	0.120
	MSE	0.040	0.020	0.070	0.060	0.050	0.030	0.000	0.030	0.030	0.030		0.030	0.030	0.060	0.030
ML2012DHS	COR	0.530	0.590	0.710	0.630	0.680	0.790	0.880	0.720	0.680	0.580	0.700	0.840	0.890	0.860	0.900
	MAE	0.110	0.120	0.190	0.190	0.120	0.140	0.070	0.060	0.150	0.080	0.080	0.110	0.110	0.120	0.050
	MSE	0.020	0.020	0.050	0.050	0.020	0.030	0.010	0.010	0.040	0.010	0.010	0.020	0.030	0.030	0.010
NG2013DHS	COR	0.610		0.820	0.810	0.850	0.860	0.920	0.790	0.770	0.600	0.850	0.890	0.910	0.850	0.860
	MAE	0.060		0.170	0.170	0.080	0.100	0.090	0.060	0.130	0.060	0.080	0.110	0.110	0.130	0.110
	MSE	0.010		0.050	0.050	0.010	0.020	0.020	0.010	0.020	0.010	0.010	0.020	0.020	0.040	0.030
NM2013DHS	COR	0.650	0.490	0.470	0.450	0.460	0.700	0.800	0.500	0.500	0.440	0.840	0.560	0.750	0.710	0.820
	MAE	0.140	0.120	0.240	0.220	0.180	0.110	0.060	0.170	0.180	0.140	0.060	0.170	0.100	0.080	0.150
	MSE	0.030	0.020	0.080	0.060	0.050	0.020	0.010	0.050	0.050	0.030	0.010	0.040	0.020	0.020	0.050
RW2015DHS	COR	0.550	0.520	0.200	0.340	0.560	0.540	0.630	0.580	0.530	0.530	0.830	0.650	0.530	0.660	0.580
	MAE	0.070	0.100	0.140	0.150	0.150	0.090	0.070	0.110	0.130	0.090	0.080	0.120	0.070	0.150	0.020
	MSE	0.010	0.020	0.020	0.030	0.030	0.010	0.010	0.020	0.030	0.010	0.010	0.020	0.010	0.050	0.000
SL2013DHS	COR	0.730	0.810	0.600	0.500	0.570	0.830	0.850	0.650	0.660	0.450	0.830	0.640	0.780	0.930	0.900
	MAE	0.100	0.080	0.180	0.170	0.140	0.120	0.090	0.080	0.140	0.080	0.070	0.110	0.140	0.090	0.070
	MSE	0.020	0.010	0.050	0.040	0.030	0.020	0.010	0.010	0.030	0.010	0.010	0.020	0.030	0.020	0.020
TG2013DHS	COR	0.680	0.730	0.520	0.620	0.690	0.810	0.850	0.630	0.610	0.580	0.710	0.790	0.840	0.940	0.870
	MAE	0.080	0.090	0.180	0.180	0.110	0.100	0.110	0.070	0.130	0.080	0.080	0.110	0.100	0.080	0.130
	MSE	0.010	0.010	0.050	0.050	0.020	0.020	0.020	0.010	0.020	0.010	0.010	0.020	0.020	0.010	0.040
ZM2013DHS	COR	0.580		0.510	0.420	0.510	0.650	0.770	0.680	0.630	0.520	0.760	0.490	0.750	0.800	0.850
	MAE	0.080		0.180	0.180	0.110	0.090	0.120	0.130	0.140	0.110	0.090	0.120	0.130	0.150	0.070
	MSE	0.010		0.050	0.050	0.020	0.010	0.020	0.030	0.030	0.020	0.010	0.020	0.030	0.050	0.020

Figure 1. Boxplots summarizing variation in performance metrics. Left column shows, for each indicator, a boxplot summarizing variation in performance metrics between surveys. Right column shows, for each survey, a boxplot summarizing variation between indicators. Rows correspond to the three performance metrics: correlation (top); Mean Absolute Error (middle); Mean Square Error (bottom). The red dashed line shows the median value for each performance metric. See Table 1 and Table 2, respectively, for full details of survey and indicator ID codes.

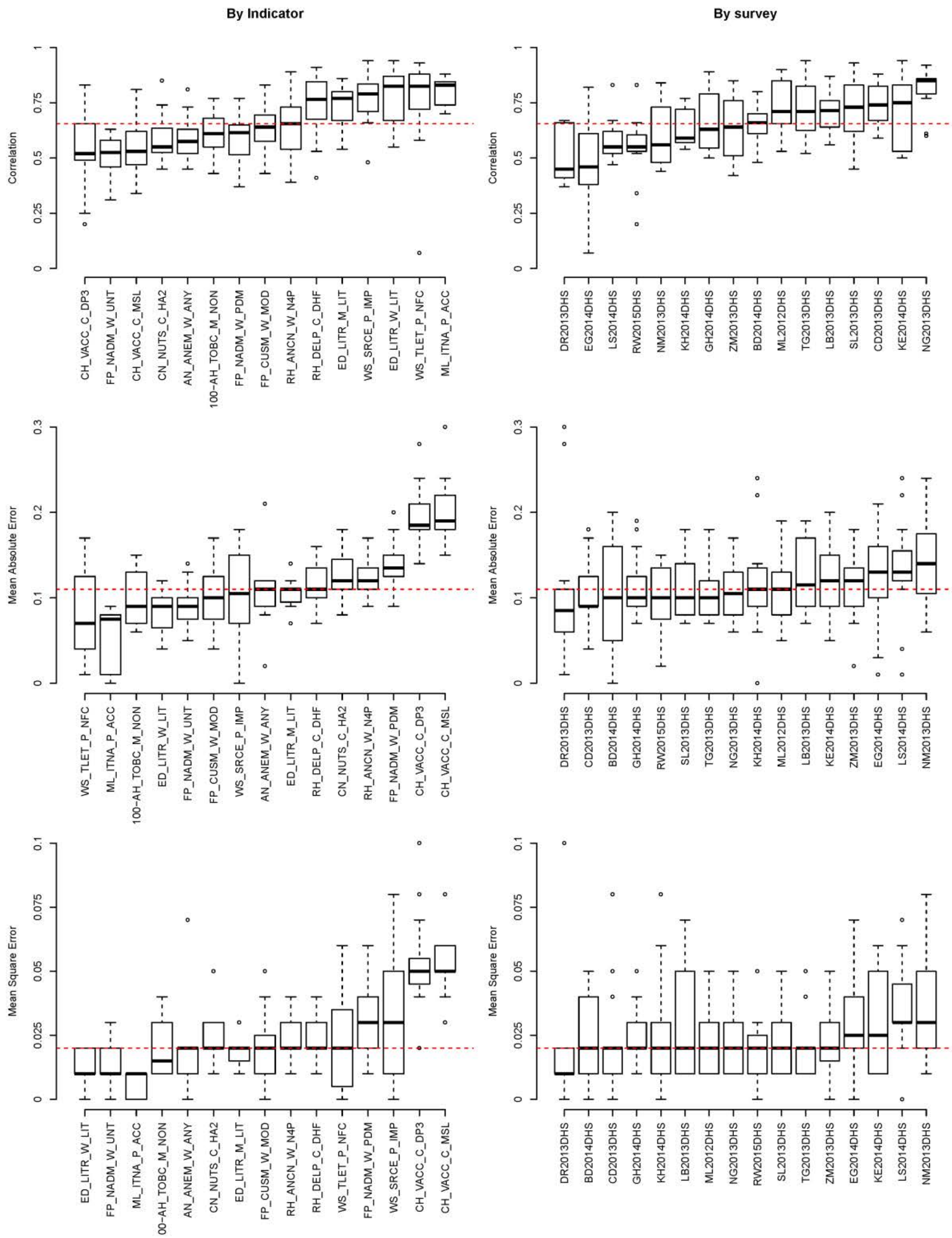


Figure 2. Out-of-sample performance (correlation) for geostatistical models predicting 15 DHS indicators (columns) for 16 DHS surveys (rows). See Table 1 and Table 2, respectively, for full details of survey and indicator ID codes. The value in each cell is the observed correlation performance for that survey-indicator pair, with the color-coding scaled according to levels of correlation (high correlation = blue, low correlation = red).

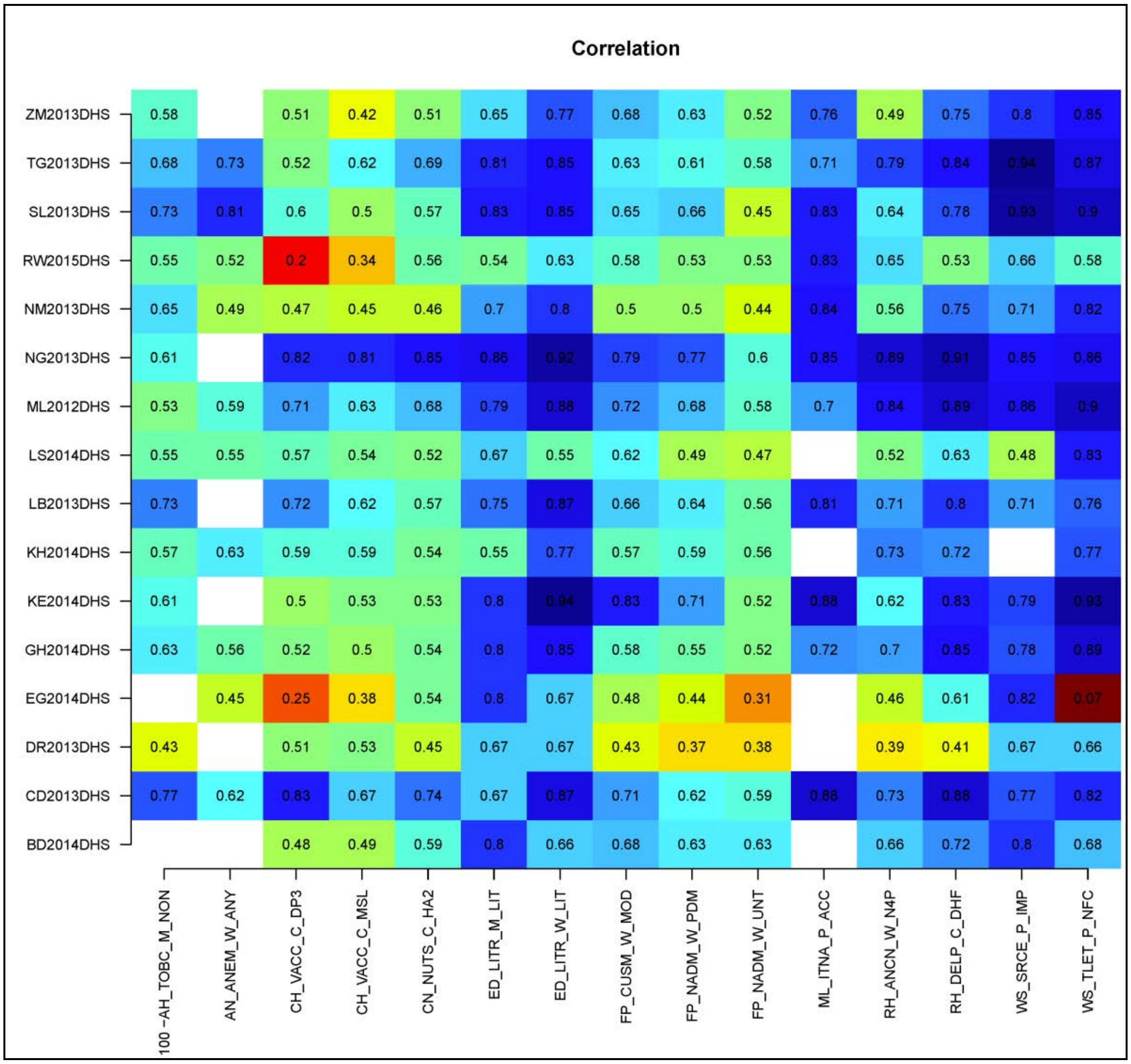


Figure 3. Out-of-sample performance (Mean Absolute Error, MAE) for geostatistical models predicting 15 DHS indicators (columns) for 16 DHS surveys (rows). See Table 1 and Table 2, respectively, for full details of survey and indicator ID codes. The value in each cell is the observed MAE performance for that survey-indicator pair, with the color-coding scaled according to levels of MAE (high MAE = red, low MAE = blue).

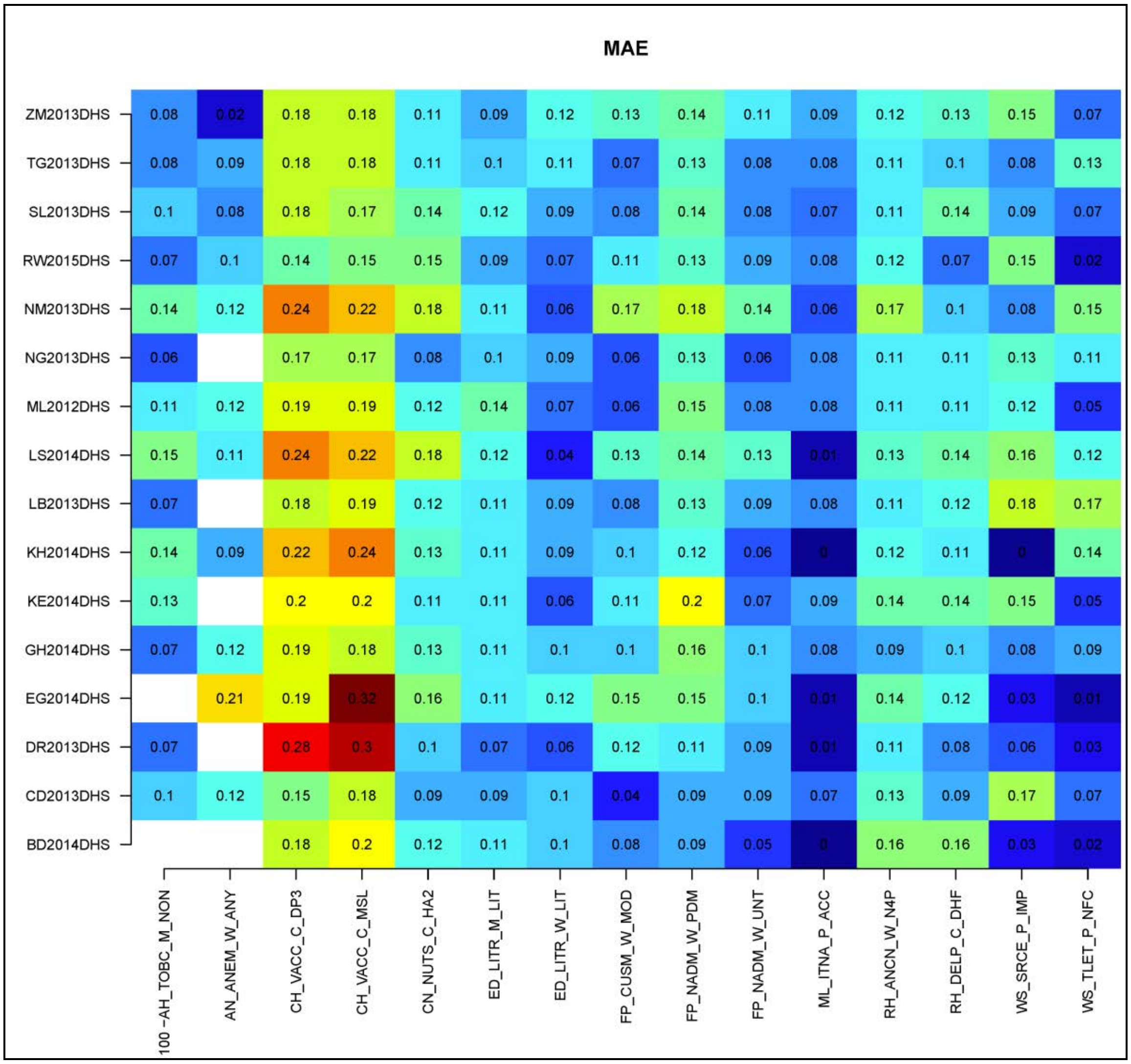
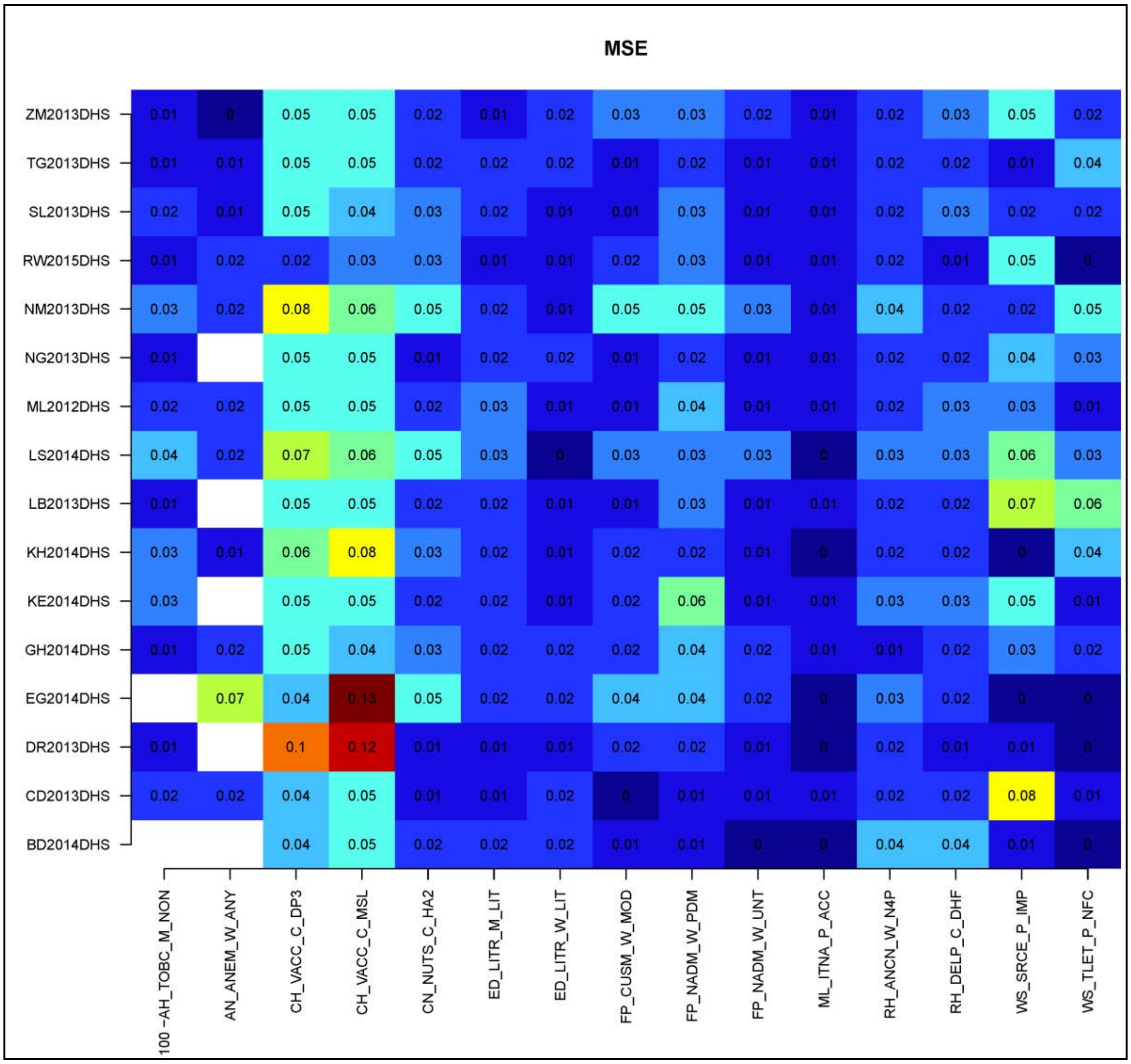


Figure 4. Out-of-sample performance (Mean Square Error, MSE) for geostatistical models predicting 15 DHS indicators (columns) for 16 DHS surveys (rows). See Table 1 and Table 2, respectively, for full details of survey and indicator ID codes. The value in each cell is the observed MSE performance for that survey-indicator pair, with the color-coding scaled according to levels of MSE (high MSE = red, low MSE = blue).



3.2.2 Correspondence between model performance and explanatory factors

Figure 5 and Table 4 give the results of the bivariate analysis to quantify the correspondence between each of the eight attributes of the cluster-level survey data and performance (COR, MAE, MSE) of the subsequent modeled surface based on those data. Table 4 details the R^2 values associated with each. COR was most strongly associated with the magnitude (and fraction) of spatially autocorrelated variance, and with the number of individuals sampled at each cluster. Conversely, MAE and MSE were most strongly associated with the magnitude of overall (cluster-level) and non-spatial variance. Figure 5 plots show the scatter of data points (each datum is an observed performance statistic for a given country-indicator modeled surface) overlaid with a fitted polynomial regression model. All relationships with higher R2 values tended to be close to linear (for example spatial variance versus COR, or non-spatial variance versus MAE and MSE). The plots for N per cluster display a dual inflexion pattern for all three performance metrics, with performance improving rapidly as N increases up to a threshold of around 50 individuals per cluster, before declining as N reaches round 150. This may be a confounding effect whereby very few clusters have numbers substantially greater than 50, and those that do may over-represent certain country-surveys and/or indicators that happen to be associated with worse performance.

Table 4. Bivariate correspondence between DHS indicator data characteristics and out-of-sample performance statistics (correlation, COR; Mean Absolute Error, MAE; Mean Square Error, MSE). Correspondence was assessed using R^2 associated with fit of polynomial model regression predicting each performance statistic as a polynomial function of each data characteristic.

Cluster-level survey data attribute	R^2 with COR	R^2 with MAE	R^2 with MSE
Point density (per km^2)	0.158	0.013	0.007
N per cluster	0.328	0.384	0.250
Cluster-level mean	0.109	0.19	0.117
Cluster-level variance	0.171	0.438	0.479
Spatial variance	0.493	0.098	0.141
Range of spatial correlation	0.081	0.017	0.016
Non-spatial variance	0.020	0.636	0.599
Spatial variance fraction	0.430	0.033	0.008

Figure 5 (part 1). Relationship between DHS indicator data characteristics (*x*-axes) and out-of-sample performance of geostatistical models generating predicted surfaces of those indicators (*y*-axes). Data characteristics assessed were density of survey clusters (1st row); mean number of individuals sampled per cluster (2nd row); cluster-level mean indicator value (3rd row); and cluster-level indicator variance (4th row). Model performance metrics assessed were correlation (left column); Mean Absolute Error (MAE, center column); and Mean Square Error (MSE, right column). Shown in each panel are the plotted values from each of 233 survey-indicators (grey dots) along with a fitted polynomial regression model and resulting R^2 value.

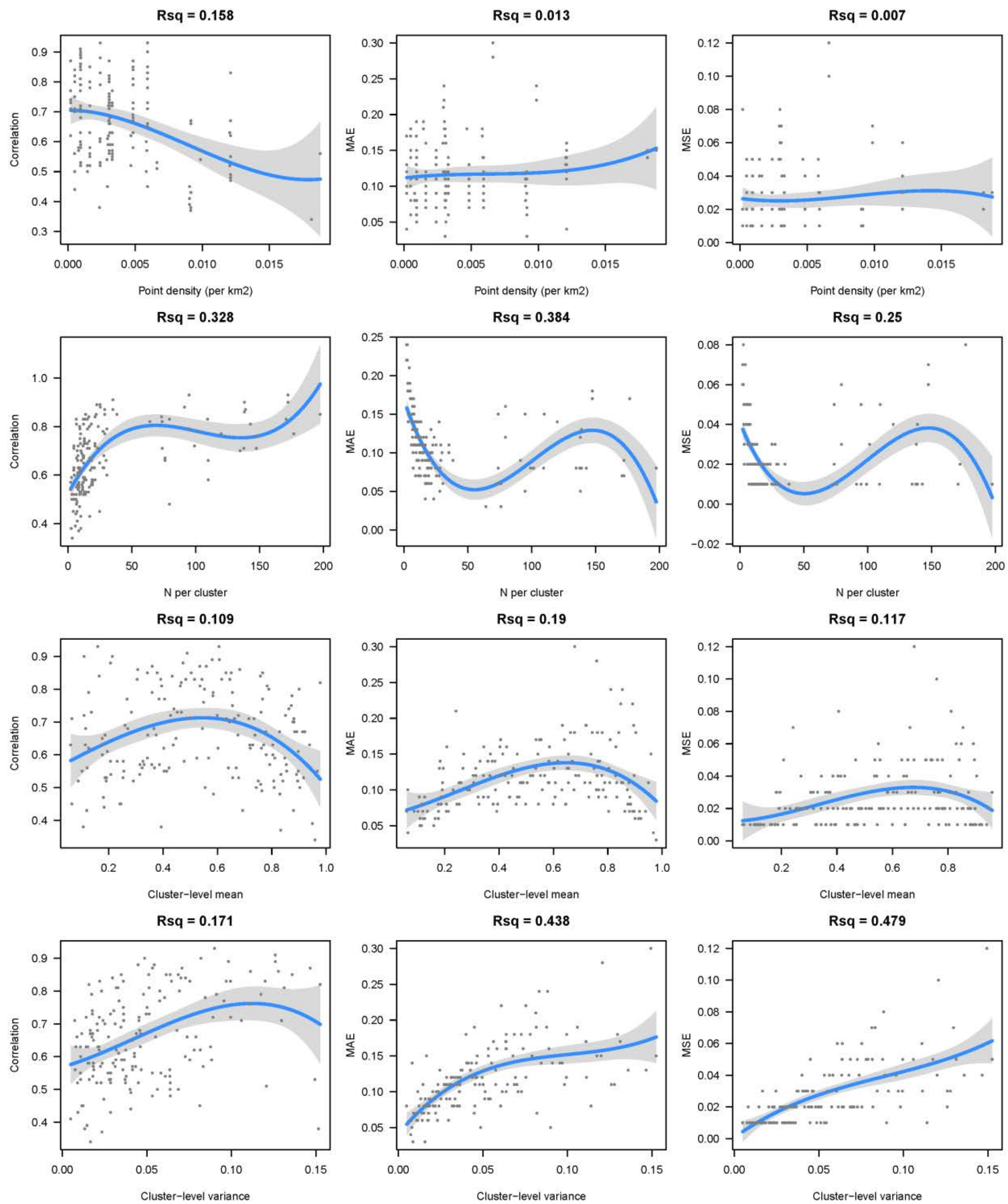
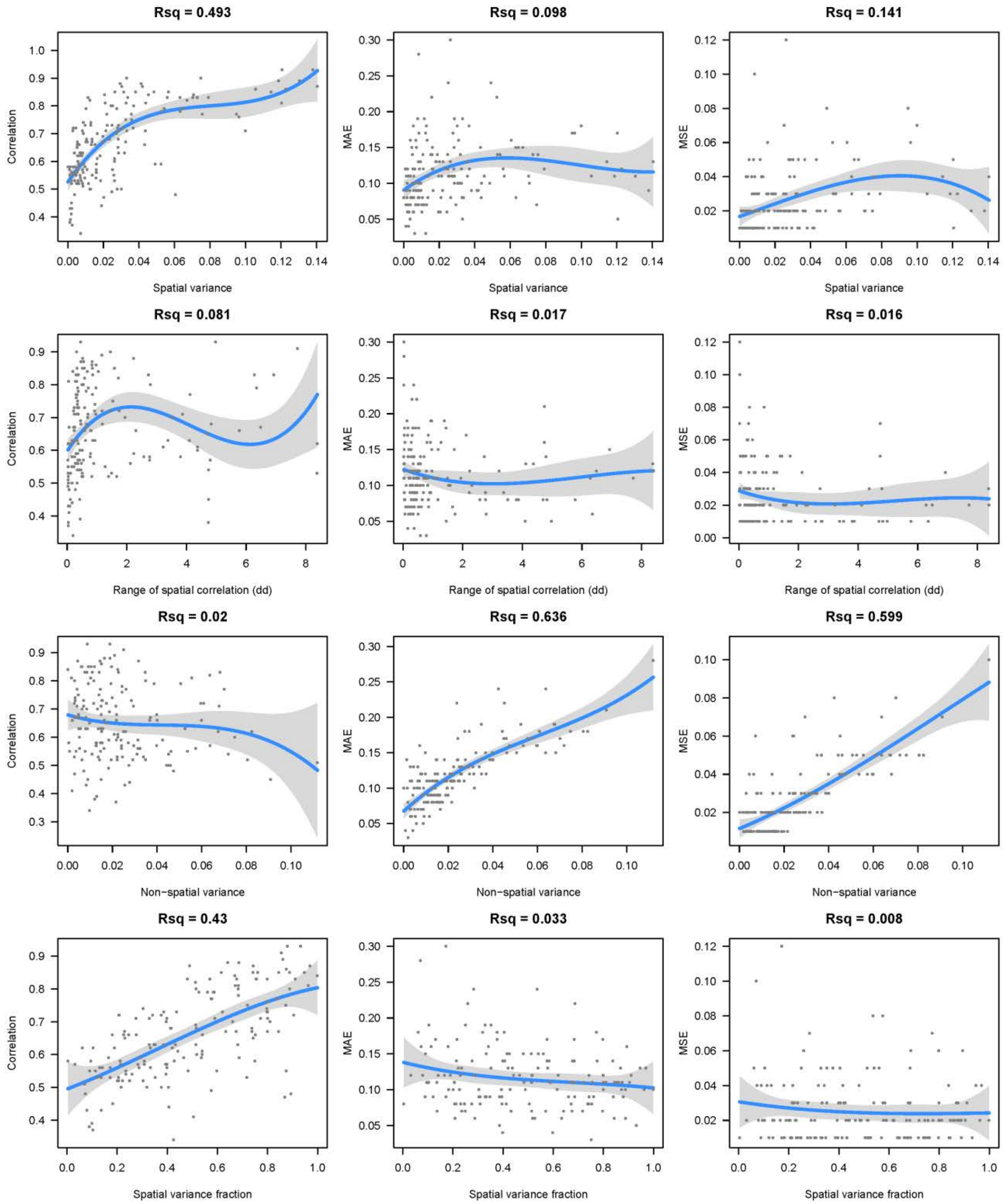


Figure 5 (part 2). See part 1 for caption.



3.2.3 Meta-modeling to predict performance a priori

Table 5 shows the results from the multivariate modeling exercise designed to test the potential to predict modeled surface accuracy a priori, that is, in advance of conducting the geostatistical modeling, based only on knowledge of the eight survey data attributes. The R2 values are fairly similar between the three performance metrics, ranging from 0.73 to 0.77. This means that about three-quarters of the variation in COR, MAE, and MSE between modeled surfaces for different indicators and different country-surveys can be explained by these eight data attributes. Consistent with the bivariate exploration just described, the individual attribute playing the largest role within the multivariate models was the non-spatial variance, which was significant at the 99.9% level in all three models. Interestingly, the second most important attribute was the cluster-level mean, which was significant at the 99% level in all three models, despite having a relatively low R^2 in the bivariate analyses.

Table 5. Coefficients from multivariate regression predicting out-of-sample performance of geostatistical models as a function of response data characteristics. Separate models were built to predict performance according to three separate out-of-sample metrics: Correlation, Mean Absolute Error (MAE), and Mean Square Error (MSE). Significance codes: 0 '*'; 0.001 '**'; 0.01 '*'; 0.05 '°'**

	Correlation			MAE			MSE		
Cluster-level survey data attribute	Estimate	P-value	Signif.	Estimate	P-value	Signif.	Estimate	P-value	Signif.
Point density (per km^2)	0.261	0.063	**	0.279	0.043	°	0.278	0.046	*
N per cluster	-0.000	0.683		-0.000	0.546		-0.000	0.636	
Cluster-level mean	0.008	0.003	**	0.007	0.011	**	0.008	0.006	**
Cluster-level variance	0.252	0.014	*	0.202	0.056	*	0.154	0.159	
Spatial variance	-0.087	0.357		-0.026	0.788		0.012	0.903	
Range of spatial correlation	-0.001	0.009	°	-0.001	0.007	*	-0.002	0.001	**
Non-spatial variance	0.506	0.000	***	0.524	0.000	***	0.607	0.000	***
Spatial variance fraction	0.007	0.129	°	0.006	0.223		0.007	0.169	
Out-of-sample R^2	0.769			0.733			0.756		

3.3 Discussion

The analyses presented in this section provide insight into the amount and magnitude of variation in accuracy of modeled spatial surfaces representing different DHS indicators and deriving from different DHS country-surveys. Some indicators perform consistently above average, and other consistently below. Similarly, certain country-surveys have a tendency to yield surfaces more or less accurate than other countries. In addition to describing these variations, the research has explored whether certain attributes of DHS survey data seem to be systematically associated with likely accuracy outcomes.

The overall amount of variation in cluster-level indicator values, and the extent to which this variation is spatially autocorrelated, are strong drivers of the accuracy of subsequent geostatistical maps. The amount of survey "effort" in terms of the density of survey clusters and the number of individuals surveyed within each cluster tended to be only weakly associated with modeled surface accuracy. Intuitively, these data volume metrics might be expected to strongly influence map accuracy, but the result may suggest simply that, since DHS surveys are designed to a broadly standardized sample specification, there is too little variation in these metrics across surveys and indicators to yield significant impact on variation in accuracy. It may also be a saturation effect, i.e., that the density of data and respondents is sufficient in most surveys to yield reasonably accurate maps, and variations around those levels have little further effect.

The results of the multivariate analyses are particularly interesting. The large R^2 values indicate that the accuracy of any given modeled DHS indicator surface, generated using the standardized geostatistical procedure described, is relatively predictable a priori based only on easily computed attributes of the pre-modeled cluster-level data. This means that, when deciding which indicators to potentially subject to a geostatistical mapping exercise, it should be possible to predict reasonably well whether the resulting maps will or will not meet a particular level of accuracy, and thus whether the exercise may or may not be worthwhile. Further, when a survey is to be prospectively designed to facilitate geostatistical spatial modeling and create modeled surfaces, the importance of different data attributes may be considered and used to influence that design.

4 Understanding Utility of Spatial Modeling for Generating Indicators at Second-Level Subnational Administrative Levels

This section describes the analysis conducted to address Objective 2, as presented in section 1.2—to explore accuracy when using spatial modeling to generate aggregate indicator estimates for second-level administrative levels, below the standard first-level units reported in DHS surveys. Methods are described first, followed by results and discussion.

4.1 Methods

Traditional DHS survey design, analysis, and presentation of results has been oriented around provision of estimates at the national level and at the first subnational unit (SNU1) level, typically called provinces, states, or regions. The geospatial work undertaken by DHS in recent years has enabled the creation of fine-scale maps providing indicator predictions across 5x5 km pixelated grids for selected countries. These maps represent the finest level of detail that is considered appropriate to predict using current data and geostatistical techniques (with some limitations on even higher resolution maps being imposed by, among other factors, the DHS geo-randomization procedure to anonymize cluster coordinate data).

This geospatial work has been motivated in part by an increasing call from the international development community and national stakeholders for reliable subnational information on demographic and health indictors to support program planning and delivery at levels more granular than SNU1. In addition to the very fine-scale pixel-level maps, therefore, there is also a remit to generate indicator estimates that represent aggregated values (e.g., indicator means) at the next lowest administrative level (e.g. SNU2, often called districts) at which detailed subnational planning is increasingly carried out. Standard DHS surveys are powered to yield pre-specified precision in indicator estimates computed directly from the survey (using appropriate survey weights) at the SNU1 level. The work described in this section was conceived to investigate the potential for using the geostatistical modeling framework to generate sufficiently precise aggregate indicator estimates at lower administrative levels, with the completion of the 2014 Kenya DHS offering a unique opportunity to address this issue.

4.1.1 The 2014 Kenya DHS

The organization of administrative tiers in Kenya was, prior to 2013, similar to many other countries across the world: eight provinces (SNU1), subdivided into 46 districts (SNU2), which were further subdivided into 262 divisions (SNU3). Following the adoption of the 2010 Kenyan constitution, however, this system was substantially revised as part of efforts to facilitate a process of devolved government. The SNU1 provinces were abolished and the 46 districts previously representing SNU2 were reconfigured to 47 counties (including Nairobi), and promoted to SNU1. These reorganizations meant that, when the 2014 DHS survey was designed, a decision was taken to power the survey to yield estimates at the new SNU1 (i.e., county) level. In practice, this meant a dramatic increase in survey effort compared both with most other DHS countries and with the preceding Kenya DHS survey conducted in 2008, which used the earlier administrative structure. The 2014 survey incorporated 40,300 households compared with 9,936 in the 2008 survey.

4.1.2 Overview of approach

The design of the 2014 Kenya DHS provides a unique opportunity to test the ability of geostatistical modeled surfaces to generate sufficiently accurate indicator estimates at what, in nearly every country other than Kenya, would be considered the SNU2 level. The approach proceeded in four main steps.

The geospatial model was used to generate County-level indicator estimates based on the full 2014 Kenya DHS data set, and these were compared with the directly calculated (i.e. weighted mean) estimates presented in the 2014 Kenya DHS report. For later reference, pixel-level modeled surfaces were also produced for each indicator using the full data set.

The full 2014 Kenya DHS data set was then artificially thinned by progressive intervals, with the most thinned version replicating the survey design of a standard DHS—based on the 2008 Kenya DHS.

The thinned data sets were then used to generate both pixel-level and County-level geospatial estimates and the effects of thinning on performance were measured.

The relative accuracy of county-level estimates derived from thinned data sets was compared between the geospatial model and directly calculated (weighted mean) estimates.

These methodological steps are now described in more detail.

4.1.3 Generation of indicator estimates at County level with appropriate uncertainty (based on full survey)

In the first step, the functionality of the standardized geostatistical framework for generating modeled DHS indicator surfaces was extended to allow "joint simulation"—the computational procedure necessary to provide correct uncertainty intervals for those estimates. For a description of the theory and computational challenges associated with geostatistical joint simulation, see Gething et al. (2010). This extended model framework was then used, initially, with the full Kenya 2014 survey to generate County-level estimates for 12 DHS indicators (see Table 2 for definitions). The purpose of this initial set of estimates was twofold: first, to allow comparison between County-level estimates based on geostatistical models versus direct calculation of indicators and confidence intervals using the weighted survey data and standard DHS formulas for standard errors; second, to provide a set of "gold-standard" estimates, based on all available data in the County-powered 2014 Kenya DHS. All subsequent estimates based on thinned sets were compared against these "gold-standard" estimates, as described next.

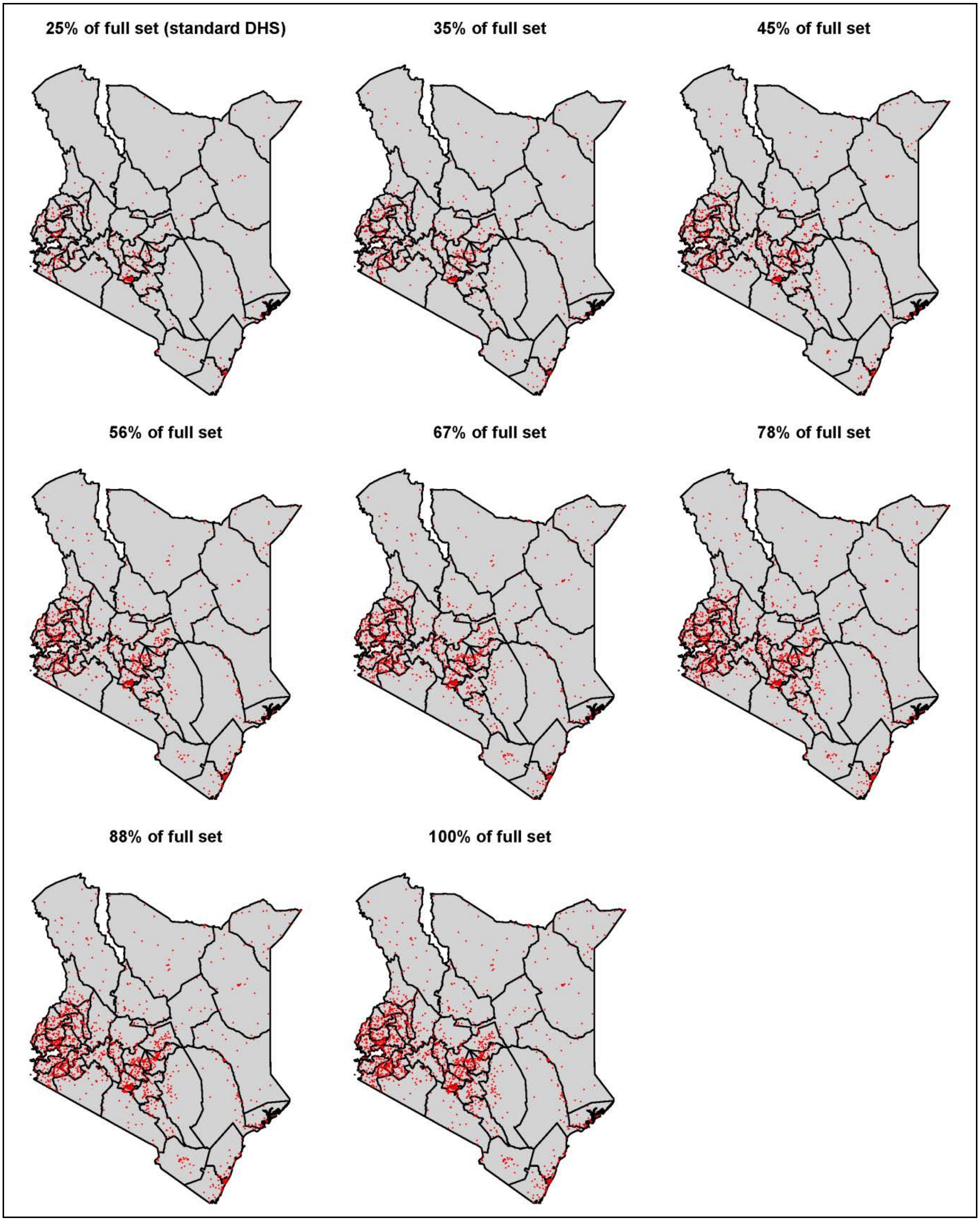
4.1.4 Creating randomly thinned survey data

In the second step, artificially thinned versions of the 2014 Kenya DHS survey were generated to form the basis for subsequent analyses. First, a thinned version was generated to emulate as closely as possible the characteristics of a standard DHS survey. Because the 2008 Kenya DHS survey was implemented prior to the 2013 administrative reorganization, it provided a template for such a design. The 2008 survey sampled 398 clusters across the country, compared with 1,594 in the 2014 survey. As such, an artificially thinned version of the 2014 survey containing just 24% of its full complement of clusters would replicate the sample size of the 2008 survey. Rather than remove the surplus 76% of points at random nationwide, it was important to emulate as closely as possible the subnational distribution and structure of the 2008 survey, which was designed within sampling strata as per standard DHS sampling procedures. The 2008 survey used five sampling strata: a rural and urban stratum for each of the seven original Kenyan provinces, plus a single urban stratum for Nairobi. The number of clusters in the 2008 survey within each of the five strata is shown in the "24%" thinning column in Table 6. To create equivalently sampled versions of the 2014 dataset, the full 2014 survey was reduced by randomly removing clusters within each stratum to achieve the required reduction to match the 2008 cluster numbers. For subsequent experimentation, it was important to minimize artefactual outcomes linked to the random selection of points in any one draw. As such, the process was repeated 100 times to yield 100 versions of the artificially thinned 2014 dataset. To explore the impact of thinning along a continuum, six further intermediate levels of thinning were defined and the process repeated for each. In total then, 100 versions of the 2014 survey were defined, artificially thinned to 24%, 35%, 45%, 56%, 67%, 78% and 88% of the complete survey. Table 6 presents the full specifications of these sets, and an example of a single draw for each thinning level is mapped in Figure 6.

Table 6. Number of survey clusters by DHS strata under progressive levels of artificial thinning relative to the full Kenya 2014 survey (100%). Thinning of 24% represented a survey equivalent to the 2008 Kenya DHS.

Stratum (based on 2008 DHS design)	Thinning level (% of full survey)							
	24%	35%	45%	56%	67%	78%	88%	100%
Central-R	41	50	60	70	79	89	99	109
Central-U	8	16	24	33	41	50	58	67
Coast-R	23	34	45	57	68	80	91	103
Coast-U	27	36	46	55	65	74	84	94
Eastern-R	49	67	85	103	122	140	158	177
Eastern-U	5	17	29	41	54	66	78	91
Nairobi-U	52	52	53	53	54	54	55	56
Northeastern-R	21	26	31	36	42	47	52	58
Northeastern-U	6	10	14	18	22	26	30	34
Nyanza-R	45	57	70	83	95	108	121	134
Nyanza-U	13	21	30	38	47	55	64	73
Rift valley-R	50	85	121	157	193	229	265	301
Rift valley-U	10	31	52	73	94	115	136	158
Western-R	37	45	53	61	70	78	86	95
Western-U	11	15	20	25	29	34	39	44
TOTAL	398	562	733	903	1,075	1,245	1,416	1,594

Figure 6. Examples of randomly thinned survey cluster sets at seven levels of thinning.



4.1.5 Evaluating performance with randomly thinned survey data

In the third step, for each of the 100 randomly thinned sets at the seven thinning levels, the geostatistical model was implemented to generate, for each indicator, a 5x5 km pixel-level modeled surface. Because this was now created using geostatistical joint simulation, it was straightforward to compute a population-weighted average within each County. This operation involved a weighted mean operation across all pixels where weights were proportional to the number of people living in each 5x5 km pixel, as defined using the WorldPop population grid for Kenya (www.worldpop.org.uk). The following two paragraphs describe the evaluation of performance of these two levels of modeled estimate using the thinned data: pixel level and County level.

4.1.5.1 Evaluating performance with thinned survey data: pixel level

The accuracy of pixel-level indicator estimates based on thinned surveys was assessed using the same out-of-sample validation procedure described in section 3.1.2.1, i.e., against cluster-level observations directly form the survey, randomly removed prior to fitting each model and held aside for reference to the predicted values at each location. By definition, it was only possible to validate the pixel-level predictions up to the 88% thinning level (i.e., not at zero thinning, the 100% level), since the latter level leaves no clusters available for hold-out. For each indicator and at each thinning level, the COR, MAE, and MSE performance statistics were computed for each of the 100 draws, along with the mean across the draws. Plots were then generated showing how the performance metrics varied as a function of thinning level.

4.1.5.2 Evaluating performance with thinned survey data: County level

Assessing the accuracy of geospatial County-level indicator estimates required a different approach than the one used for the pixel-level estimates described above. Although pixel-level estimates can be compared directly with held-out cluster-level observations from the survey, there are no directly observed County-level observations to compare against. Instead, there are two different approaches that can act as a gold-standard estimate against which to compare the estimates derived from thinned sets. One option is to use the directly calculated DHS indicator point estimates that are calculated as a simple weighted mean of the survey observations within each County, with weights reflecting population weighting, adjustments for non-response, and other factors. An alternative option is to consider as the gold-standard reference the geospatial County-level estimates made using the full 2014 survey, as described in section 4.1.3.

This distinction is an important one for many of the subsequent results in this section because, as described in more detail later, the geospatially derived versus directly calculated County-level estimates often differ by a non-negligible amount, and so the choice of one or the other as a gold-standard reference will inevitably heavily influence the resulting performance statistics. The relative merits of each approach are discussed in more detail subsequently, but neither is clearly superior. As such, all subsequent assessments of performance were carried out in duplicate: one set using the geospatially derived full-survey County-level estimates as the gold standard, and another using the directly calculated estimates.

As with the pixel-level tests, County-level performance was evaluated for each indicator, at each thinning level, and over 100 random draws, generating the COR, MAE, and MSE performance statistics for each. Plots were then generated showing how the performance metrics varied as a function of thinning level using both alternative gold-standard references.

4.1.6 Comparison of geospatial versus directly calculated (weighted mean) indicator accuracy

As well as investigating the impact of progressively smaller survey sample sizes on the performance of spatial modeled surfaces in absolute terms, a practical question concerns how this performance compares in relative terms with the default approach of directly computing weighted means of the survey data. This question becomes particularly pertinent if the goal of estimating indicators accurately at SNU2, rather than SNU1, becomes more generally pursued in future DHS and other nationally representative surveys. One approach for ensuring adequate accuracy at SNU2 level is, of course, to dramatically increase sample size, as was done for the 2014 Kenya DHS. This is inevitably laborious and expensive. A key question, then, is whether the use of geostatistical estimation would allow equivalent accuracy at SNU2 level without the need for such a large increase in sampling effort.

To directly enumerate the potential added value of using geostatistical models to generate SNU2 estimates, the concept of "equivalent survey size" is introduced. This approach attempts to enumerate how much larger a survey would need to be when used in conjunction with a default weighted mean indicator calculation approach if it were to match the accuracy provided by a standard-sized survey in concert with a geostatistical model. Any additional survey size required by the former approach to match the performance of the latter can then be interpreted as the added value of the geostatistical model—since it has achieved the same result with the lesser sampling effort.

Our study implemented this approach as follows: First, County-level estimates were made using the geostatistical model based on the fully thinned survey draws (with only 24% of the full complement of clusters in each draw, and replicating the sample size of the standard 2008 Kenya DHS survey) and the performance of these estimates evaluated against the gold-standard references. Second, weighted mean County-level estimates were made based on usual DHS weighted mean calculations, but using thinned surveys (every level of thinning between 24% and 88%, as specified in Table 6). Again, performance was evaluated against both gold-standard references. The "equivalent survey size" was calculated by identifying the level of thinning that yielded DHS weighted mean calculations with performance equivalent to the geostatistical estimates based on the 24% (fully thinned) survey. This procedure was carried out for all indicators and with the usual 100 draws for each thinning level.

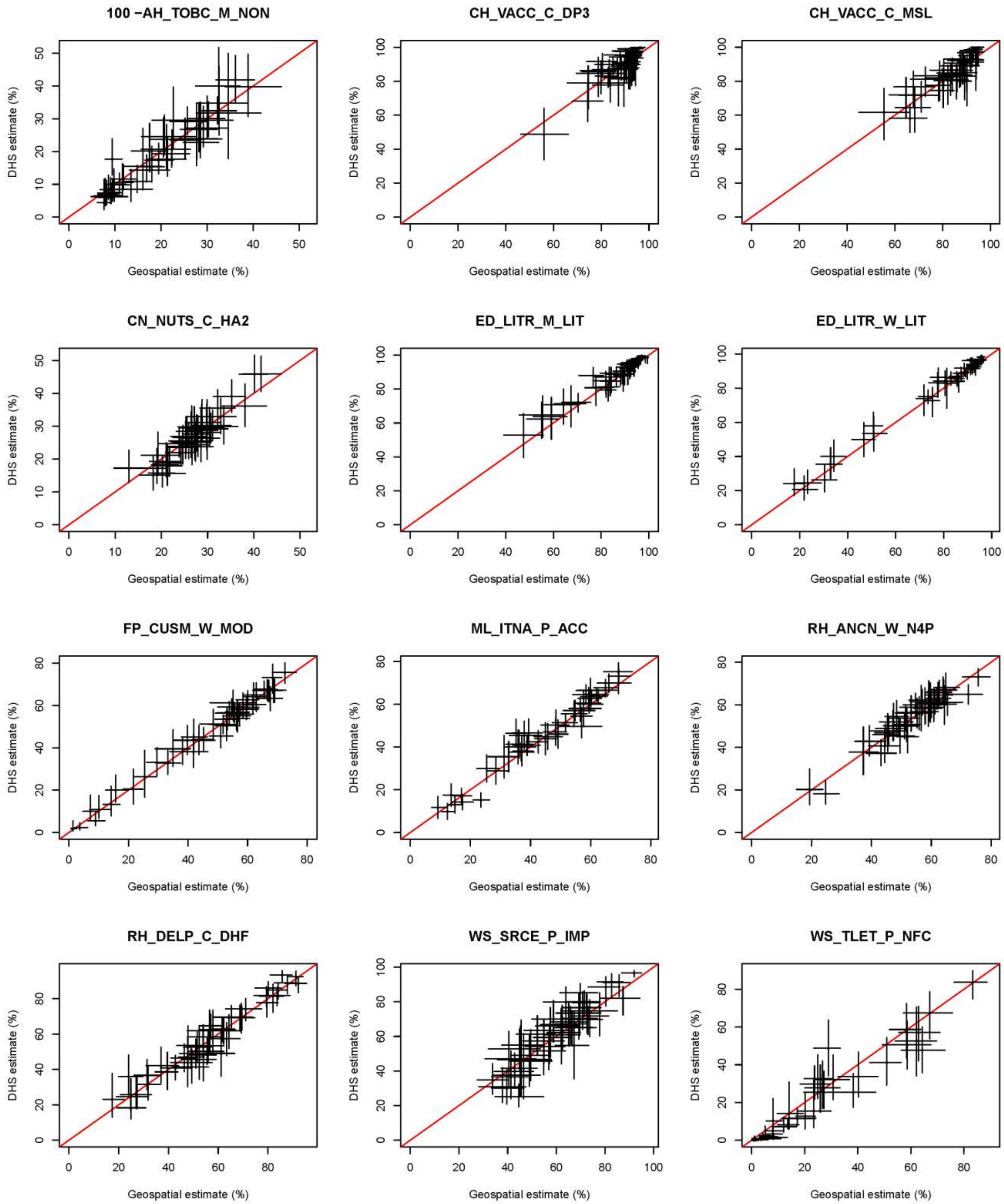
4.2 Results

4.2.1 Comparison of County-level geostatistical indicator estimates versus directly calculated (weighted mean) indicators, based on the full 2014 Kenya DHS survey

Figure 7 visualizes the correspondence between the estimates of County-level indicators based on the full Kenya 2014 survey, using the default DHS weighted-mean-of-survey-data calculation and standard error formulas (*y*-axes), and the equivalent County-level estimates made using the geostatistical model (*x*-axes). Each plotted cross represents the pair of estimates for a single Kenyan County. As Figure 7 shows, most crosses cluster tightly around the identity (i.e., 1:1) line. Although for some indicators there is noticeable deviation from the 1:1 line when assessing point estimates (i.e., the center of each cross), the uncertainty intervals fail to intersect that line only rarely—indicating that it is uncommon for the geostatistical and directly calculated estimates to differ with statistical significance.

More detailed results of the comparison between geostatistical and directly calculated County-level estimates are shown for each indicator in Annex 1, Figure 11 A-L. These additional plots include County-by-County comparisons of the two estimates, along with individual scatterplots comparing the point estimates and the upper and lower uncertainty bounds of each estimate.

Figure 7. Comparisons of County-level estimates as derived directly from 2014 Kenya DHS survey data versus from geostatistical model for each DHS indicator. Each cross represents estimates for a particular County, with length of lines representing 95% credible/confidence intervals around estimates. See Table 2 for full details of indicator ID codes.



4.2.2 *Performance of geospatial estimation with thinned survey data: pixel level*

Figure 8 shows the results of the evaluation of the impact of survey size on performance of the geostatistical model in generating pixel-level estimates (i.e., modeled surfaces). For each indicator, a plot is shown for COR, MAE, and MSE and each shows how the performance statistic changes as the survey size ranges between the full 2014 Kenya DHS (100%) and the standard DHS size (24% of the full set). At each thinning level, the performance statistic calculated from each of the 100 random draws is plotted (grey dots) along with the mean value across the 100 sets at each thinning level (black dots and line). The general tendencies displayed reflect what might be expected: correlation nearly always increases with increasing survey size, while MAE and MSE decrease, reflecting steadily improving precision and accuracy. These tendencies are more pronounced for some indicators than other, however, and the trends are not always monotonic. This latter feature potentially suggests that even averaging over 100 random realizations of each thinned set does not fully mitigate stochasticity associated with the particular set of clusters retained for each thinning level.

Figure 8 (part 1). Performance of geospatial model (correlation, left column; mean absolute error, middle column; mean square error, right column) in predicting pixel-level indicator values using progressively thinned survey sets. Each thinning level was assessed using 100 randomly thinned subsets (grey dots) and the mean across all sets is shown in solid black. Each row is a separate indicator. See Table 2 for details of Indicator ID codes.

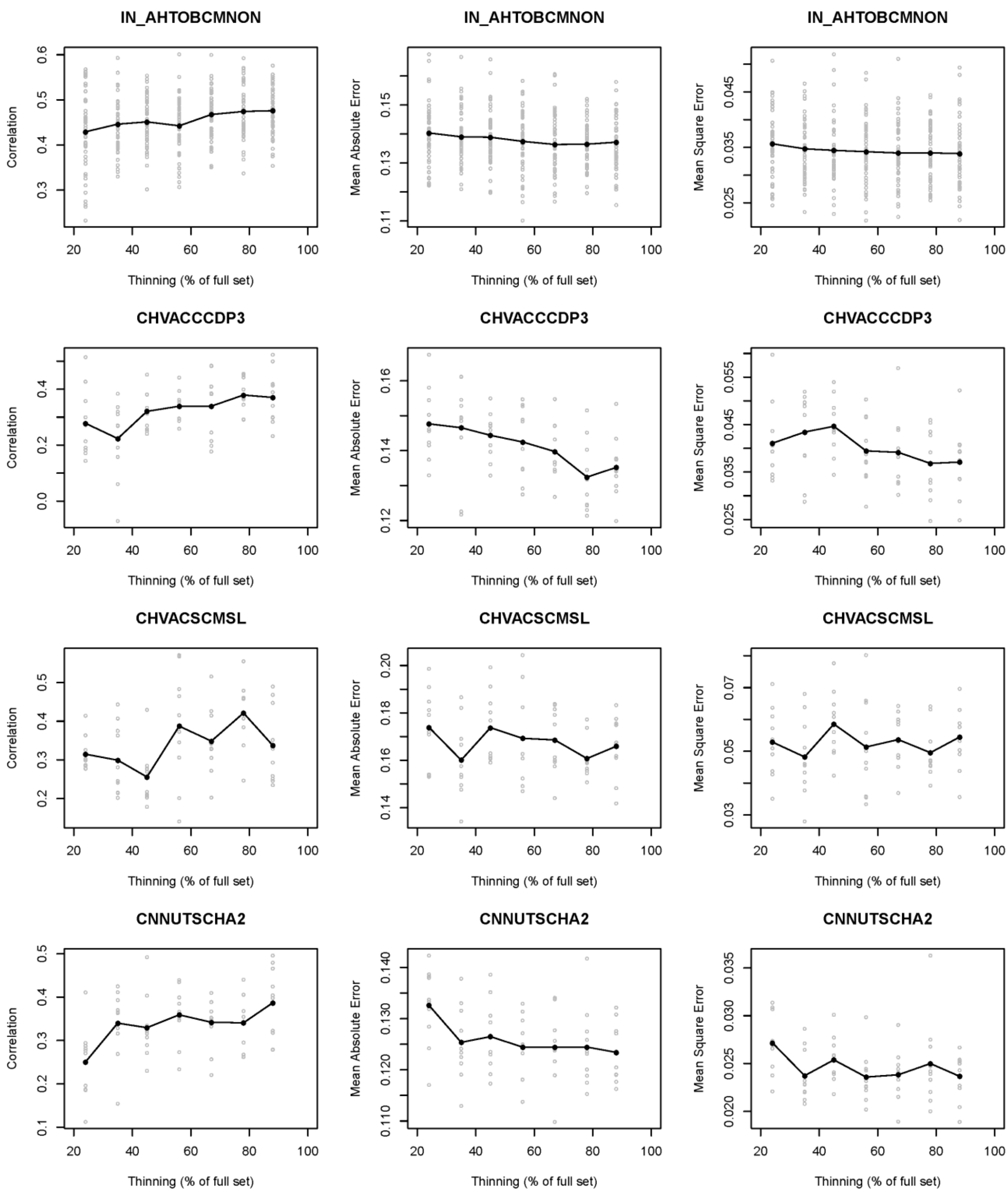


Figure 8 (part 2). See part 1 for caption.

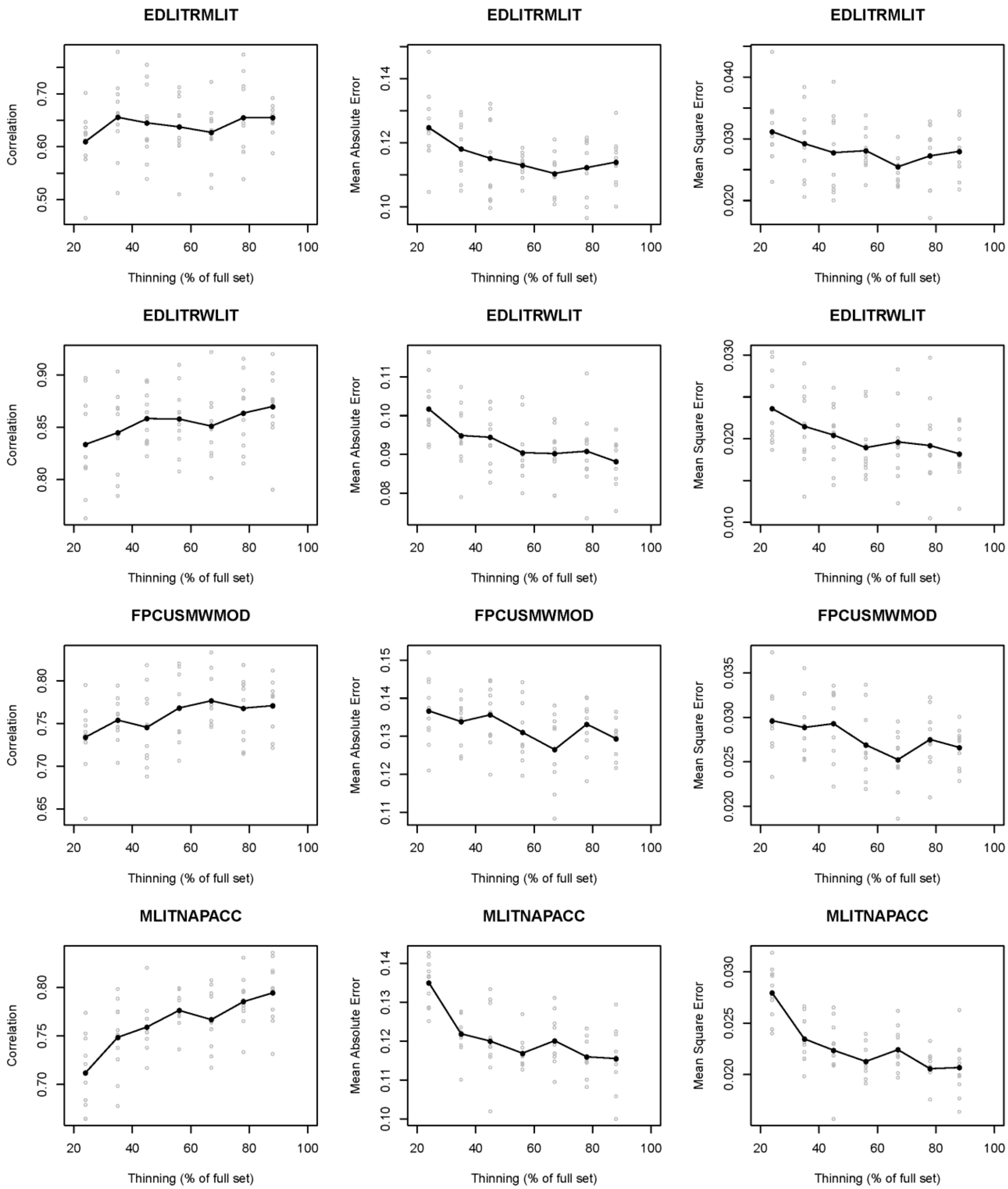
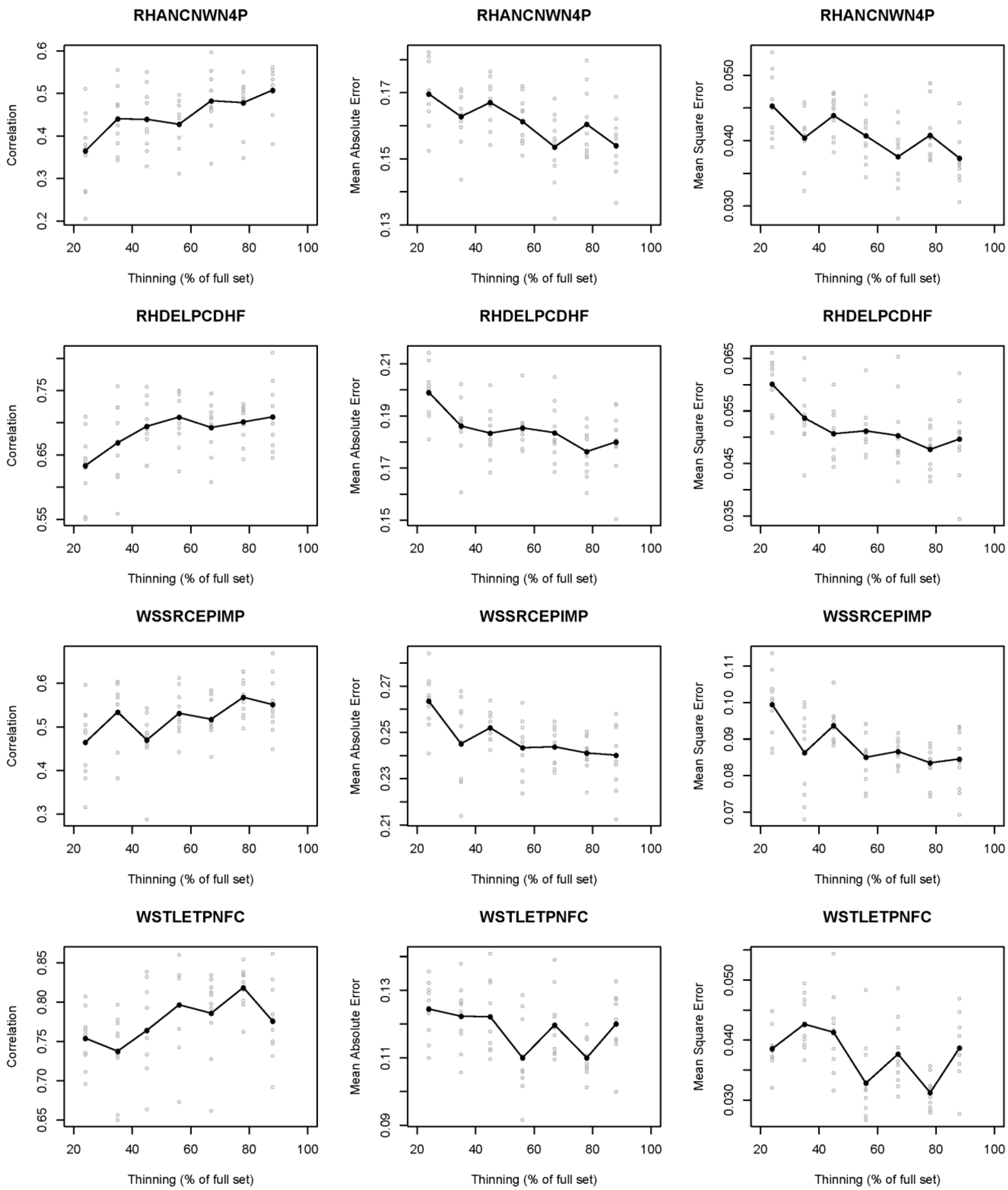


Figure 8 (part 3). See part 1 for caption.



4.2.3 Performance of geospatial estimation with thinned survey data: County level

Figure 9 shows the results of the evaluation of the impact of survey size on performance of the geostatistical model in generating County-level estimates. Again, for each indicator, a plot is shown for COR, MAE, and MSE and each shows how the performance statistic changes as the survey size ranges across the different thinning levels. Because there are two alternative gold-standard references for these County-level estimates, two sets of results are shown on each plot. Red lines denote performance against the full-survey weighted-mean-of-survey-data indicator estimates, and black lines denote performance against the full-survey geostatistical estimates. Once again, broad patterns appear as expected, with performance progressively improving with increasing survey size for all three performance metrics. Unlike with pixel-level assessment, the trends at County-level are much smoother, with correlation nearly always increasing monotonically with survey size, and MAE and MSE monotonically decreasing. Interestingly, the trends tend to be non-linear: performance gains are progressively smaller as survey size is progressively increased, such that much of the gain is achieved in moving from 25% to 50% of the full survey, with a move from 50% to 75%, or even to 100%, contributing proportionately less gain in performance.

Comparison of the red versus black lines on each plot demonstrates the importance of the choice of gold-standard reference. While use of either choice yields the same qualitative trends in performance versus survey size, the performance at any given survey size is always much better when using the geospatially-derived gold standard rather than the directly calculated weighted-mean-of-survey-data. This is not surprising since, as we have already seen in Figure 7 and Figure 11, the two candidate gold standard estimates often differ considerably for any given County. Of course, this is partly by design because the geostatistical estimates incorporate various additional or alternative sources of information compared with the directly calculated weighted means with the deliberate purpose of better capturing fine-scale variation (and thus ultimately to allow more accurate aggregate estimates). In particular, the geostatistical estimates: (1) borrow strength across County boundaries; (2) are informed by a suite of environmental and sociodemographic covariates; and (3) use a different underlying representation of population distribution when computing weighted means (gridded population rasters from WorldPop, whereas DHS survey design and weights draw on census data). Since these features are all retained in the geostatistical modeling based on thinned surveys it is expected that the geostatistical gold standard will be closer to the thinned estimates than will the directly calculated weighted means.

Figure 9 (part 1). Performance of geospatial model (correlation, left column; mean absolute error, middle column; mean square error, right column) in predicting County-level mean indicator values using progressively thinned survey sets. Performance was evaluated against two alternative gold standards: directly-calculated DHS indicator estimates (red) and geostatistical estimates based on full (unthinned) original survey (black). Each thinning level was assessed using 100 randomly thinned subsets (grey/pink dots), and the mean across all sets is shown in solid black/red. Each row is a separate indicator. See Table 2 for details of Indicator ID codes.

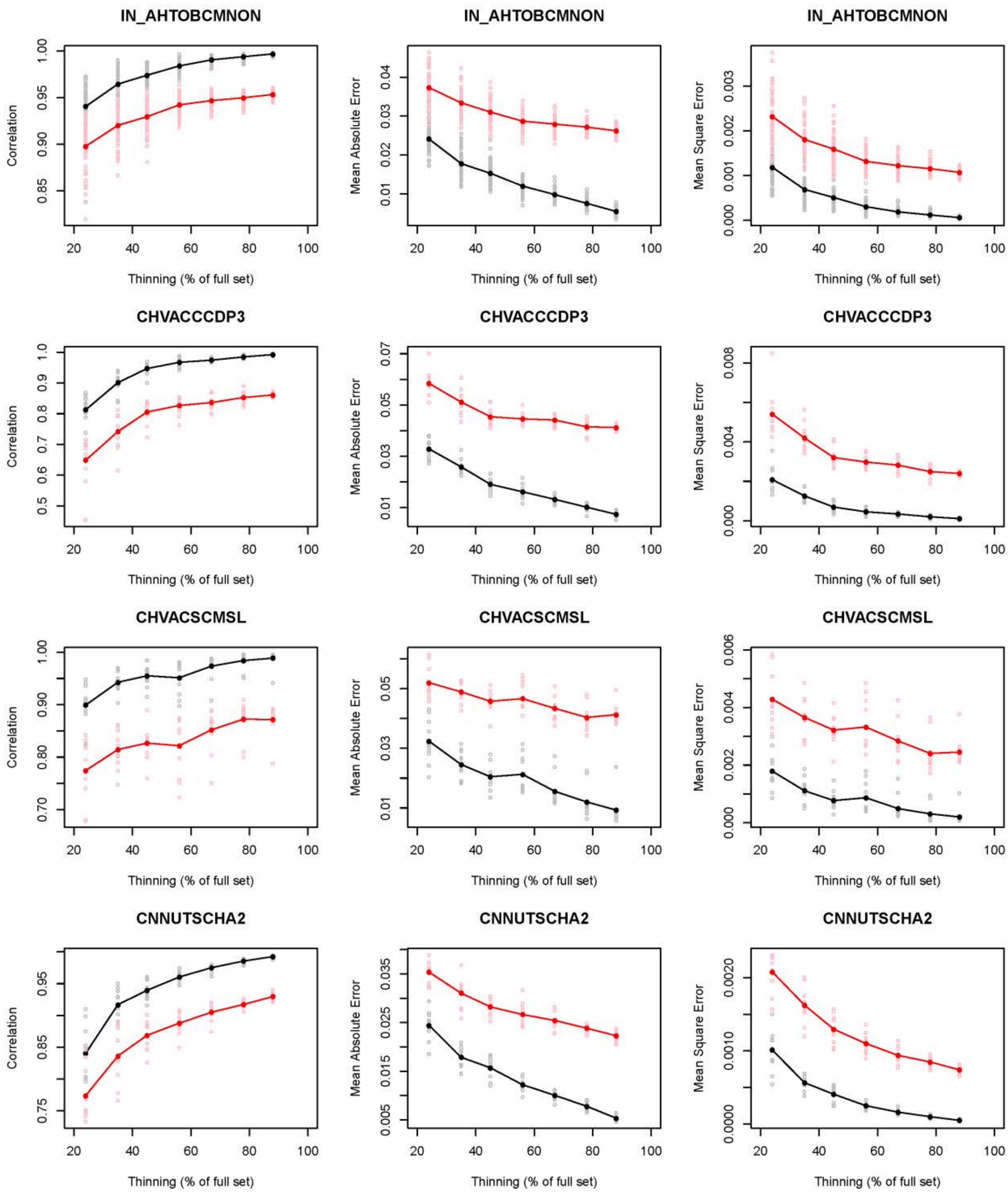


Figure 9 (part 2). See part 1 for caption.

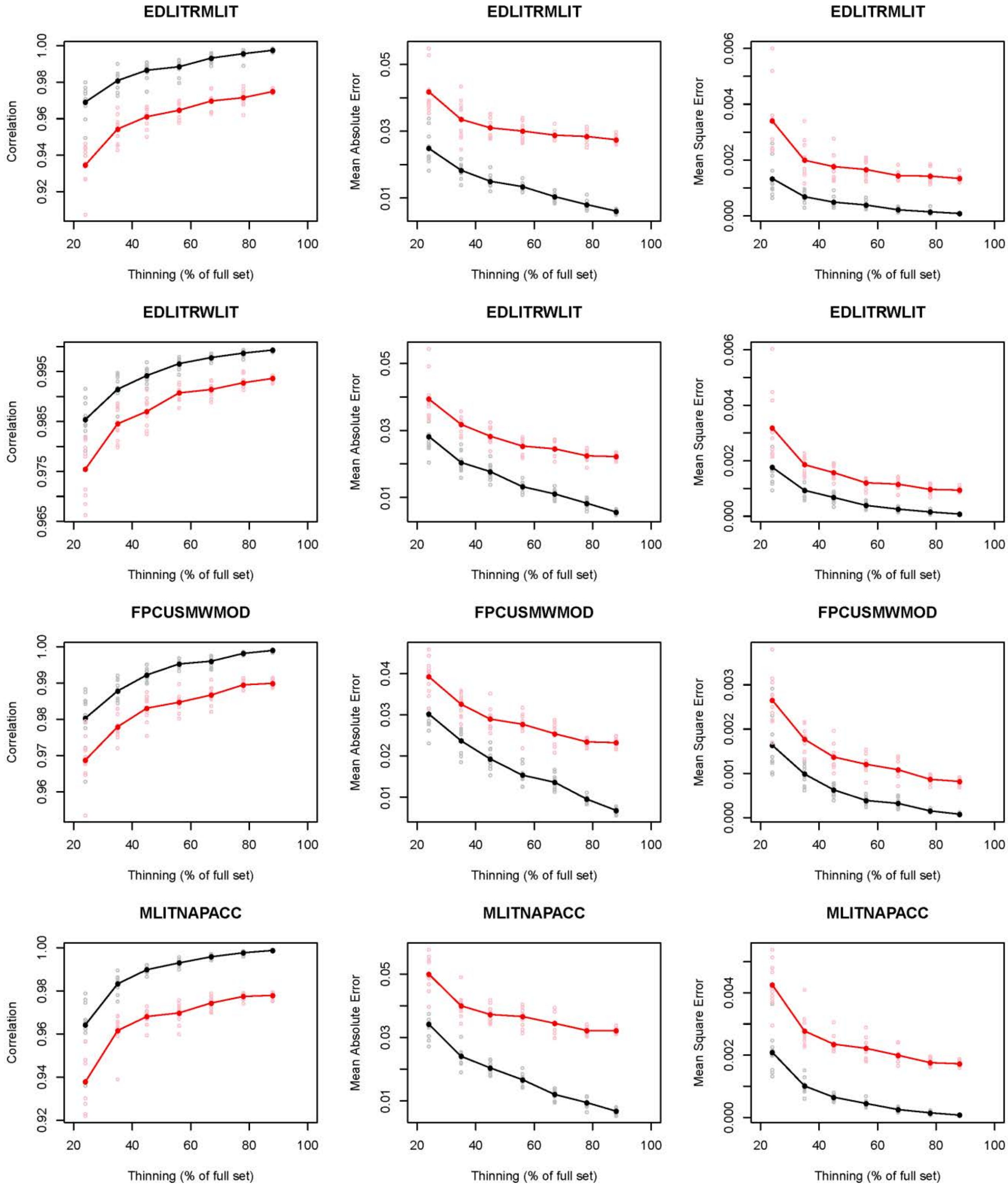
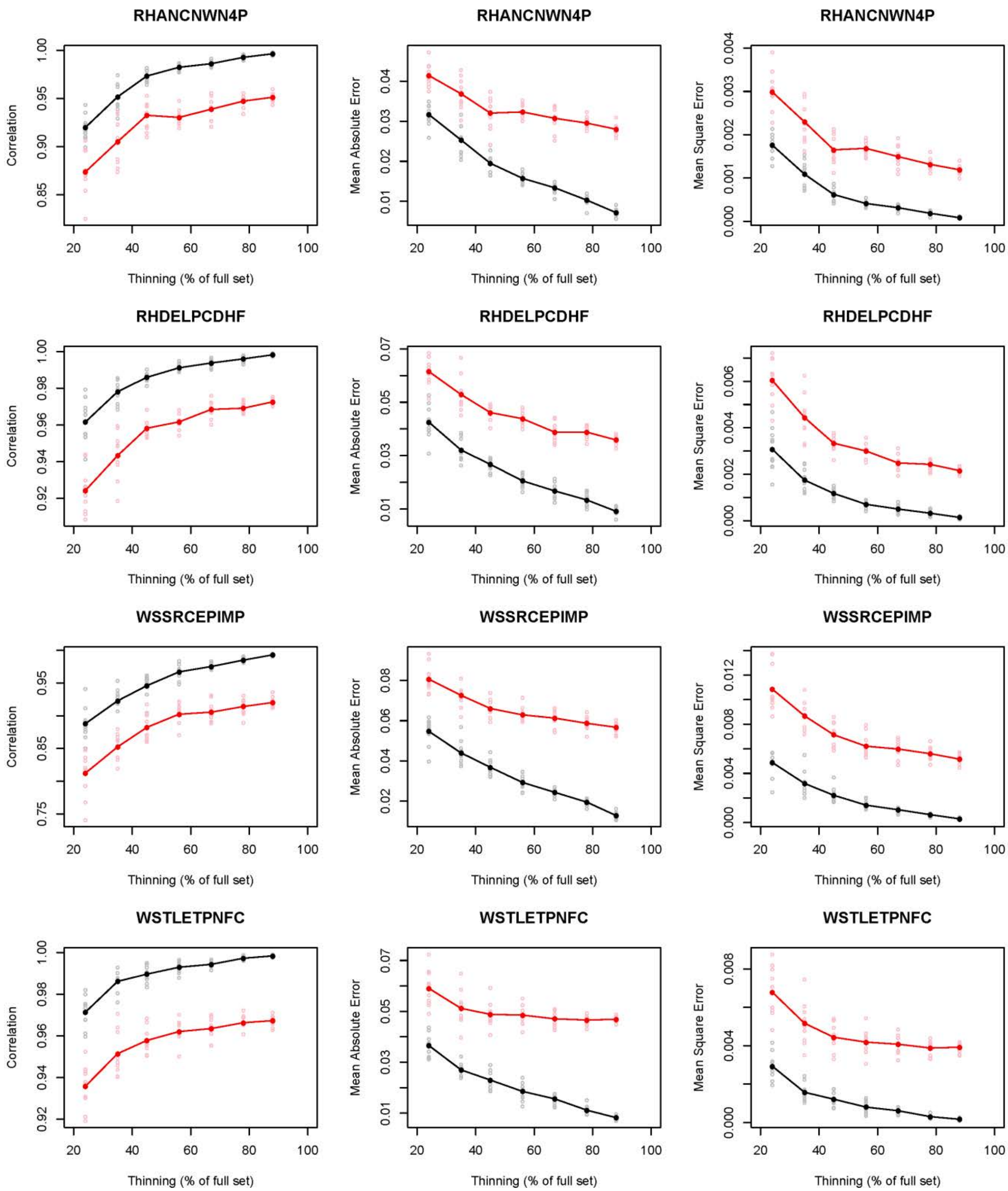
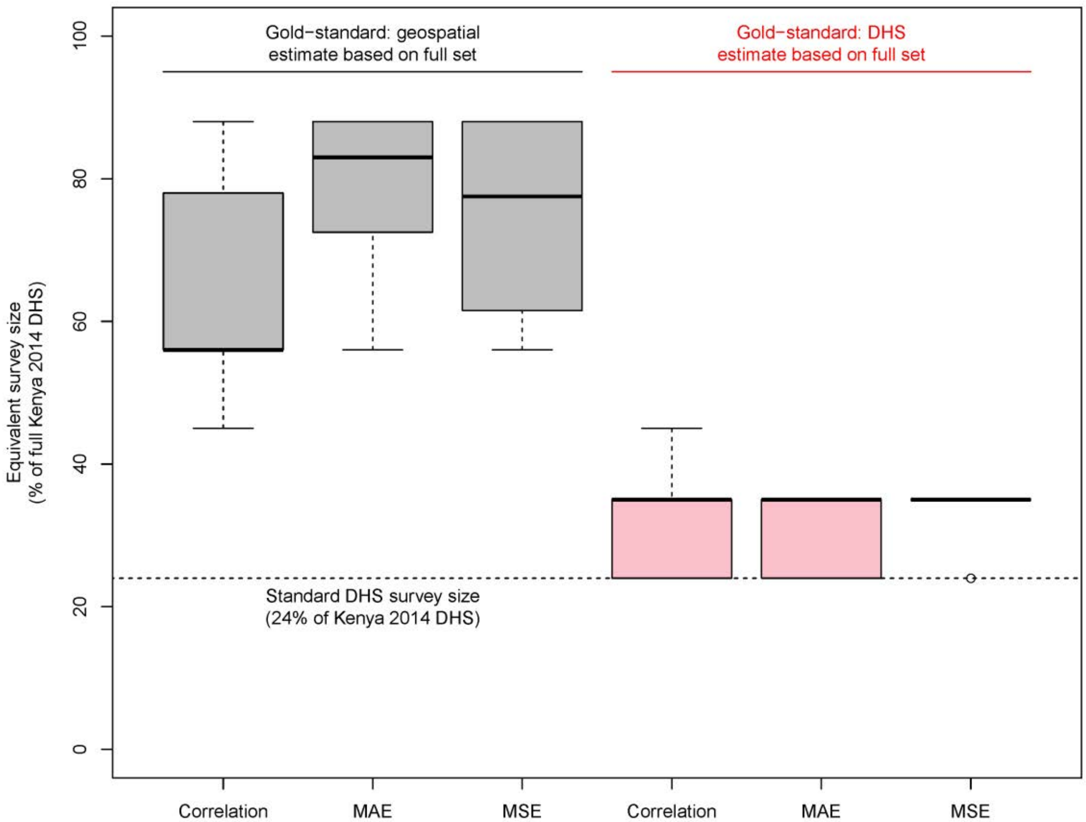


Figure 9 (part 3). See part 1 for caption.



4.2.4 Comparison of geospatial versus directly calculated (weighted mean) indicator accuracy at County level

Figure 10. Performance of geospatial model for estimating County-level mean indicator values based on standard DHS survey sizes. Geospatial model performance is expressed using the "equivalent survey size" concept, indicating the size of survey that would have yielded equivalent performance in the absence of a geospatial model. Performance was measured against two gold standards: geospatial County-level estimates based on the full survey (black/grey boxplots on left) and direct DHS indicator estimates based on the full survey (black/pink boxplots on right). Boxplots show distribution of performance values across all 12 indicators and for the three metrics: Correlation (COR), Mean Absolute Error (MAE), and Mean Square Error (MSE).



This section summarizes the results of the exercise to estimate the "equivalent survey size" yielded by using the geostatistical model with the fully thinned survey compared with larger surveys using only directly calculated weighted-means-of-survey-data. Figure 10 presents six boxplots. These represent the variation across County-level indicator estimates for each of the three performance metrics when assessed against the two alternative gold-standard references. Table provides the full breakdown of performance metrics per indicator for all 12 indicators. Again, it is clear that the results differ markedly depending on which gold standard is used. When using the geostatistical-derived gold standard, the geostatistical County-level estimates based on the standard DHS survey size (24% thinning) are approximately as accurate as a survey three times larger (64% thinning to achieve same correlation, 78% thinning to achieve same MAE, 75% thinning to achieve same MSE) using the default weighted mean approach. However, if the gold standard is switched to the DHS County-level estimates based on the full-set, the added value of geostatistical modeling is much diminished, equivalent to a survey of about 32% thinning and so adding only modest value above the 24% thinned survey on which the model was based.

Table 7. Performance of geospatial model for estimating County-level mean indicator values based on standard DHS survey sizes. Geospatial model performance is expressed using the "equivalent survey size" concept, indicating the size of survey that would have yielded equivalent performance in the absence of a geospatial model. Performance was measured against two gold standards: geospatial County-level estimates based on the full survey (left hand columns) and direct DHS indicator estimates based on the full survey (right hand columns). Three performance metrics are given: Correlation (COR), Mean Absolute Error (MAE), and Mean Square Error (MSE).

	Equivalent survey size as % of full DHS (Gold standard: geospatial estimate based on full survey)			Equivalent survey size as % of full DHS (Gold standard: DHS estimate based on full survey)		
	COR	**MAE**	**MSE**	**COR**	**MAE**	**MSE**
IN_AHTOBCMNON	78	88	88	35	35	35
CHVACCCDP3	56	88	88	24	24	24
CHVACSCMSL	88	88	88	24	35	35
CNNUTSCHA2	45	78	67	24	35	35
EDLITRMLIT	78	88	88	45	24	35
EDLITRWLIT	56	56	56	35	35	35
FPCUSMWMOD	56	56	56	35	35	35
MLITNAPACC	45	67	56	24	24	24
RHANCNWN4P	56	78	67	35	35	35
RHDELPCDHF	67	78	67	35	35	35
WSSRCEPIMP	56	88	88	24	35	35
WSTLETPNFC	88	88	88	35	24	35
Mean	**64.1**	**78.4**	**74.8**	**31.3**	**31.3**	**33.2**

4.3 Discussion

Using the unique opportunity provided by the exceptionally densely sampled 2014 Kenya DHS survey, the analyses presented in this section provide new insights about the potential utility of geostatistical models for generating aggregated indicator estimates at administrative levels below SNU1, as well as showing more broadly the impact of survey size on model performance at different levels.

First, we see that indicator estimates made at county-level are broadly similar when generated using the geostatistical model versus calculating them directly from the weighted survey data. We do, however, also see important differences. These arguably represent the fact that geostatistical models first attempt to estimate fine-scale pixel-level indicator values (taking into account information from covariates and drawing strength from data points within a radius) before these are subsequently aggregated up to the desired spatial units. Other sources of disparity will arise from the fact the geostatistical aggregate estimates draw upon a pixel-level representation of the population distribution, rather than the coarser census enumeration areas on which the DHS population-based sampling frame is designed.

Second, we see that surveys with greater sampling density than a standard DHS design yield improvements in the results of geostatistical models, whether at pixel level (i.e., the modeled surfaces) or when those surfaces are aggregated to generate mean indicator estimates across administrative regions. These improvements are more marked and more consistent in the case of the aggregated estimates than the pixel-level estimates. The rate of improvement is not linear, and most gains are made with the first doubling of cluster numbers rather than the second.

Third, we see that the use of a geostatistical model to estimate aggregated indicator estimates tends to yield more precise estimates than the default approach of directly calculating weighted means of survey data. The extent of this improvement is heavily dependent on the choice of gold standard reference. Given the theoretical advantages of the full-survey geostatistical-derived estimates described, it is reasonable to consider these comparisons as most informative. Under this rationale, we see that, on average across the indicators and performance metrics, the use of a geostatistical model based on a standard DHS survey to estimate indicators at "standard" SNU2 level (i.e., at a level of geographical aggregation equivalent to SNU2 in most countries) yield results of equivalent accuracy to a survey three times larger in the absence of a geostatistical model. This finding raises the prospect of wider use of geostatistical methods to potentially reduce the sampling effort, and thus the resource cost, of future DHS surveys.

4.3.1 Implications and recommendations

The extent to which these findings can be generalized to other countries and settings can ultimately only be verified by further quantitative analysis using new data sets. However, it seems likely that similar results would be expected, since the main factors driving the relationships explored here, namely the design of DHS surveys and the intrinsic statistical properties of the resulting indicator data, tend to be broadly similar across settings. One practical consideration, of course, is the definition of SNU2 units in different countries. In this analysis we have used the Kenya County level as a proxy that is approximately equivalent to the size of SNU2 units found in most other countries across Africa and beyond. Where this approximation is not appropriate – for example in a country with SNU2 units substantially smaller than the Kenyan County – then the rule-of-thumb results presented here would need to be modified.

An immediate and general recommendation arising from this work is that the integration of geospatial methods in the survey design and subsequent data analysis stages should be considered for future DHS surveys, especially if a desired outcome is to provide precise estimates below the SNU1 level. This integration will provide more precise estimates below the SNU1 level and do so with lesser requirements for large sample sizes.

To fully operationalize this recommendation, the analyses presented here would require a number of extensions. First, repeating the work implemented here for the 2014 Kenya DHS to at least one other survey with similarly dense sampling would provide verification of the generalizability of the findings. Second, a generalized framework could be envisioned that would allow prospective survey design to be carried out to achieve pre-specified levels of precision using the geospatial model, taking into account the exact nature of SNU2 units in a given country. Such a framework would require the current analysis to be repeated for a synthetic set of administrative units at progressively smaller levels of aggregation, to provide reference results against which actual units in other countries could be compared.

References

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. http://dhsprogram.com/pubs/pdf/SAR9/SAR9.pdf.

Burgert, C.R., T. Dontamsetti, A. Marshall, P. W. Gething. 2016. *Guidance for Use of The DHS Program Modeled Map Surfaces*. DHS Spatial Analysis Reports No. 14. Rockville, Maryland, USA: ICF International.

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. http://dhsprogram.com/pubs/pdf/SAR7/SAR7.pdf.

Diggle, P., and P. J. Ribeiro. 2007. *Model-Based Geostatistics*. Springer Science & Business Media. New York, NY.

Diggle, P. J., J. Tawn, and R. Moyeed. 1998. "Model-Based Geostatistics." *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 47(3):299-350.

Gething, P., A. Tatem, T. Bird, and C. R. Burgert-Brucker. 2015. *Creating Spatial Interpolation Surfaces with DHS Data*. DHS Spatial Analysis Reports No. 11. Rockville, Maryland, USA: ICF International. http://dhsprogram.com/pubs/pdf/SAR11/SAR11.pdf.

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

Gething, P. W., A. P. Patil, S. I. Hay. 2010. "Quantifying Aggregated Uncertainty in Plasmodium Falciparum Malaria Prevalence and Populations at Risk via Efficient Space-Time Geostatistical Joint Simulation." *PLoS Computational Biology*, e1000724.

Hay, S. I., C. A. Guerra, P. W. Gething, A. P. Patil, A. J. Tatem, A. M. Noor, C. W. Kabaria, B. H. Manh, I. R. Elyazar, S. Brooker, D. L. Smith, R. A. Moyeed, and R. W. Snow. 2009. "A World Malaria Map: Plasmodium Falciparum Endemicity in 2007." PLoS Med 6(3):e1000048.

Annex 1. Further Results: Kenya County Level Comparisons (Unthinned)

Figure 11A. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is 100-AH_TOBC_M_NON.

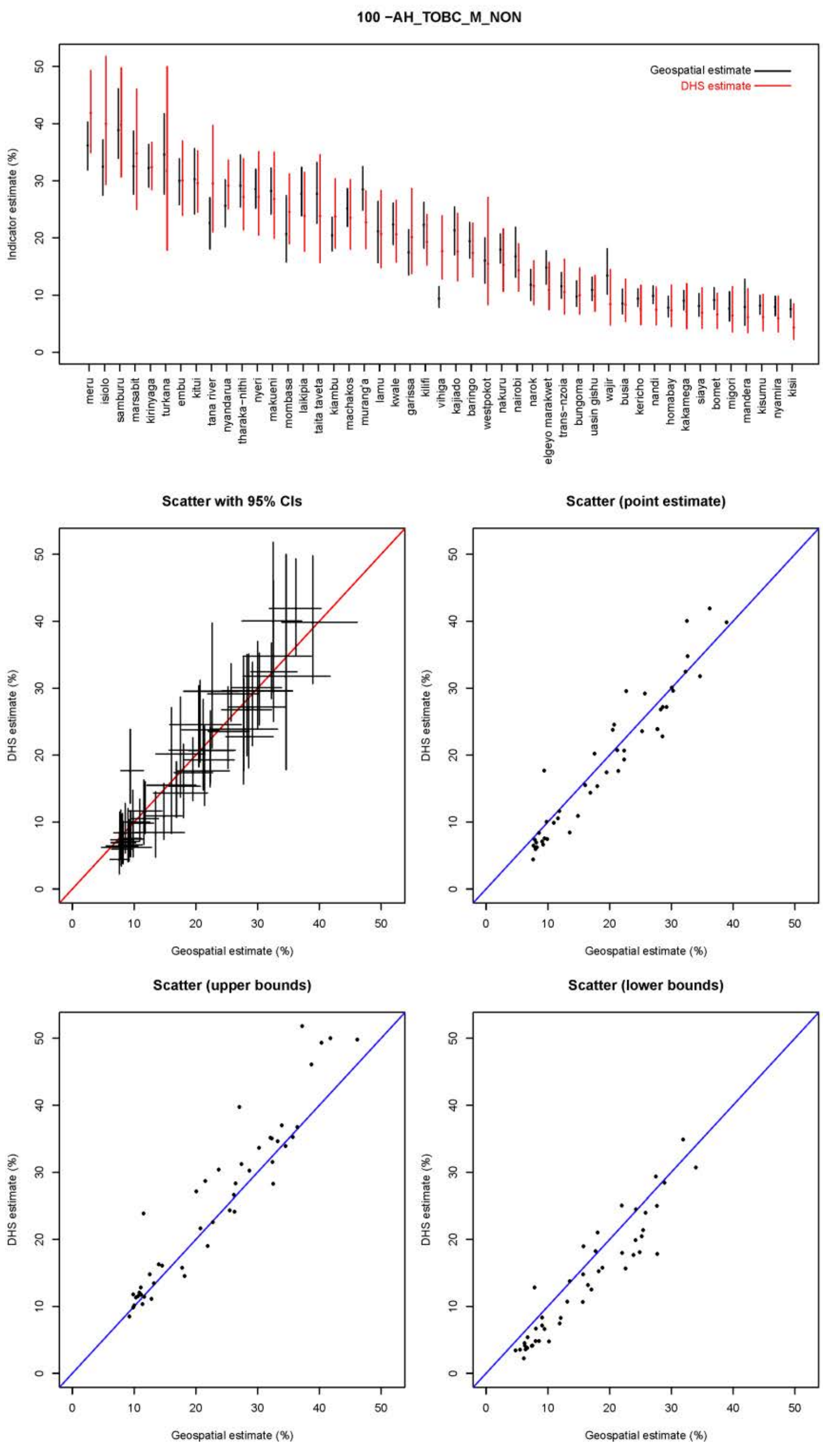


Figure 11B. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is CH_VACC_C_DP3.

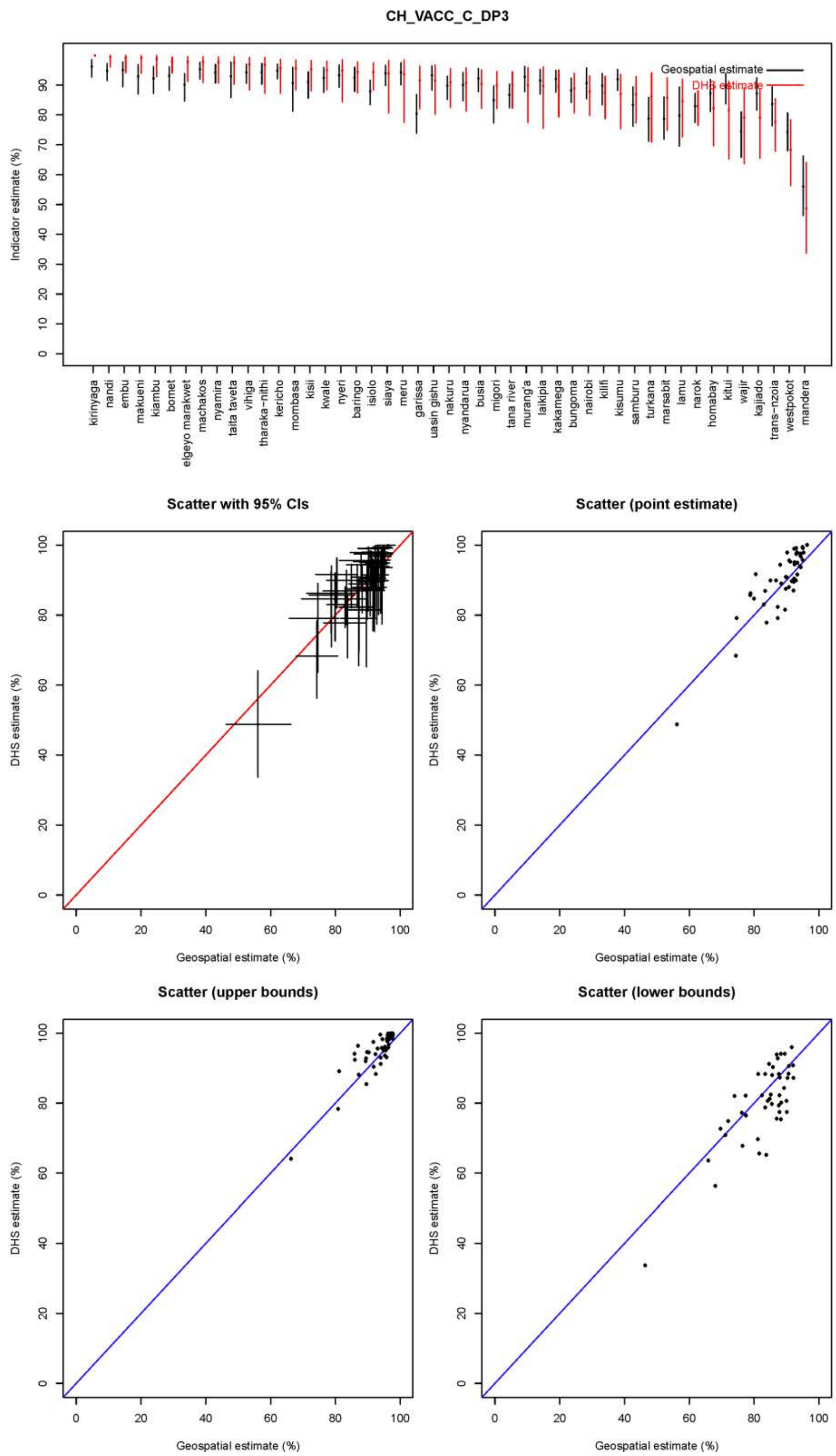


Figure 11C. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is CH_VACC_C_MSL.

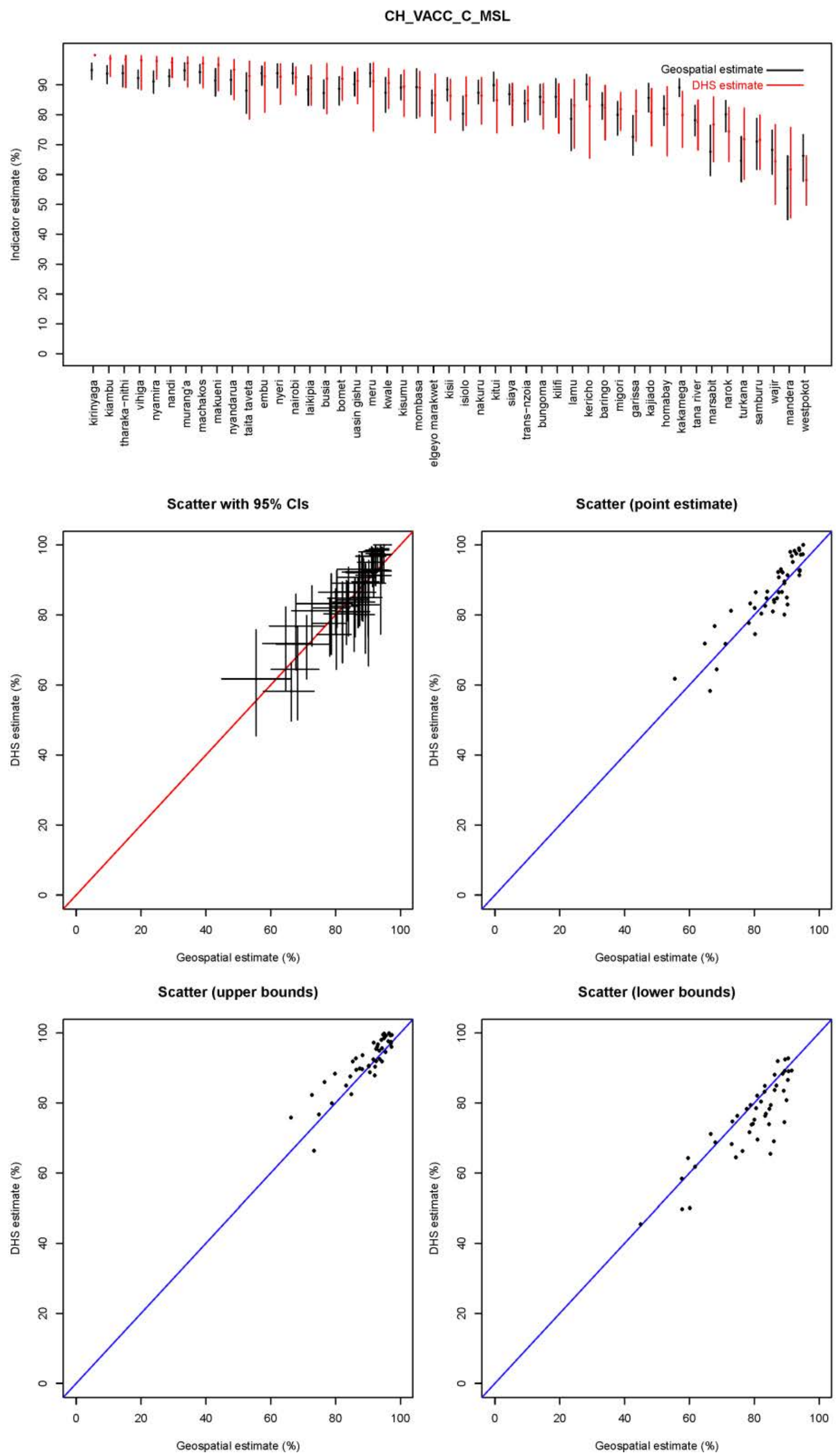


Figure 11D. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is CN_NUTS_C_HA2.

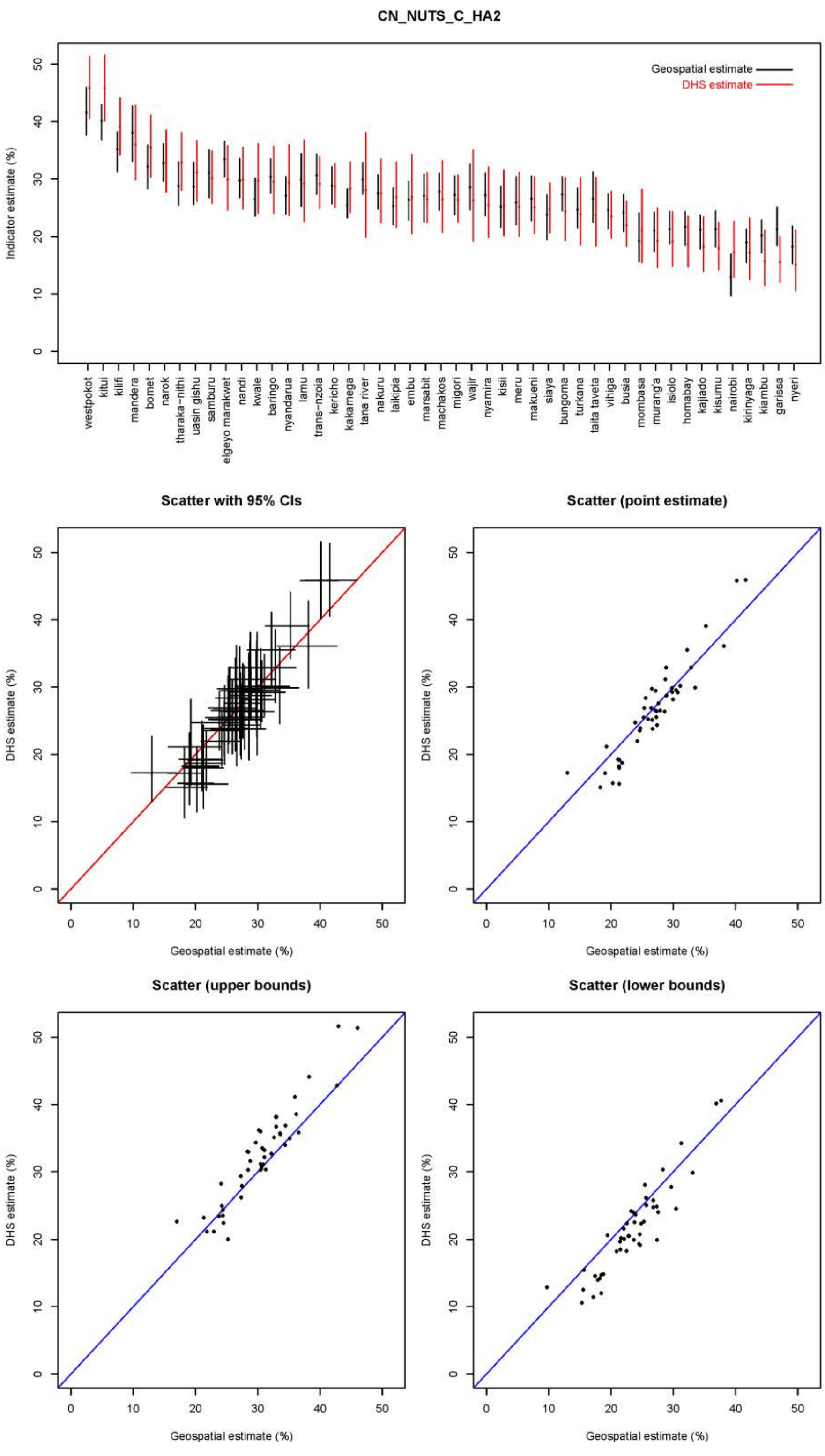


Figure 11E. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is ED_LITR_M_LIT.

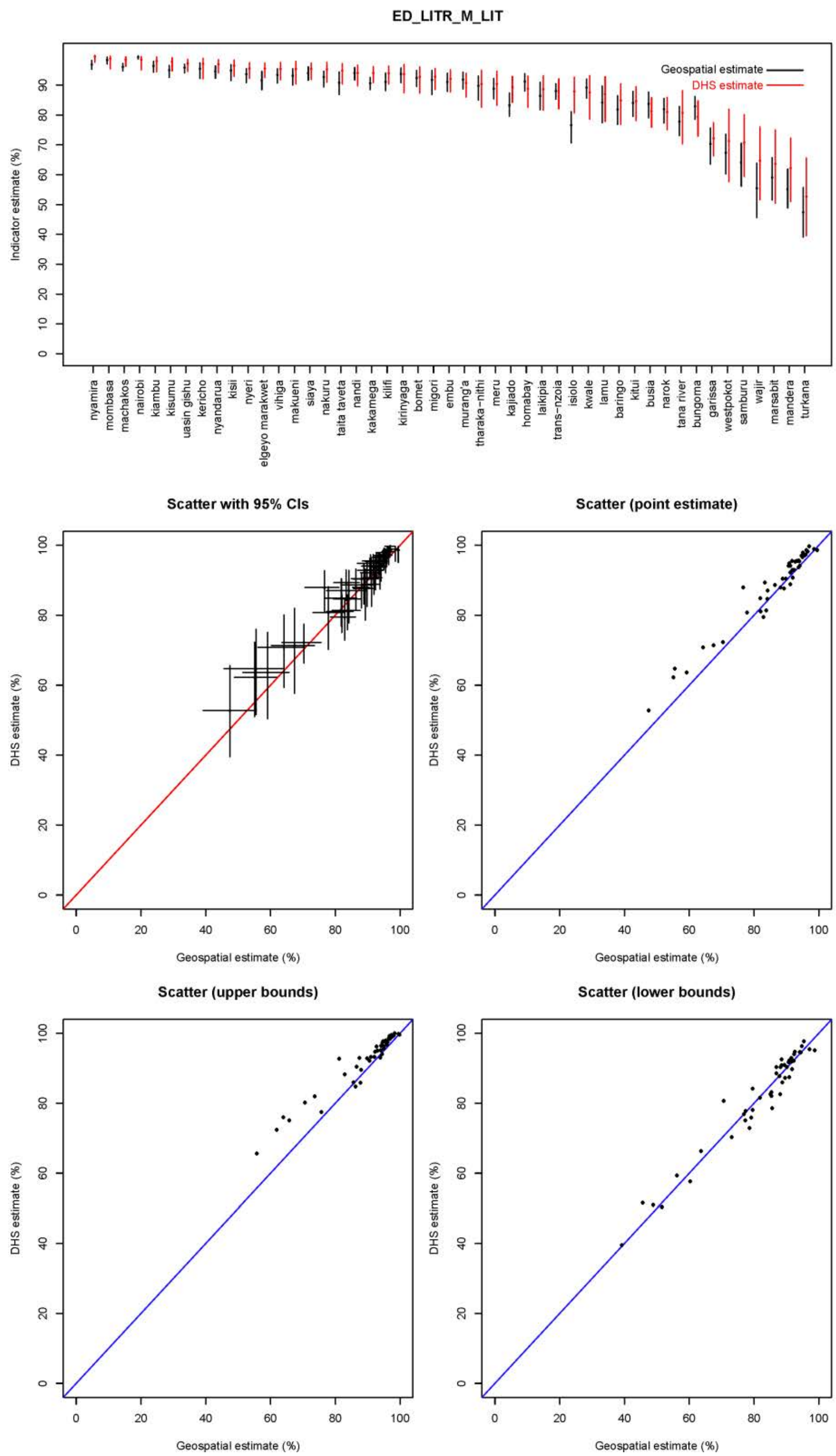


Figure 11F. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is ED_LITR_W_LIT.

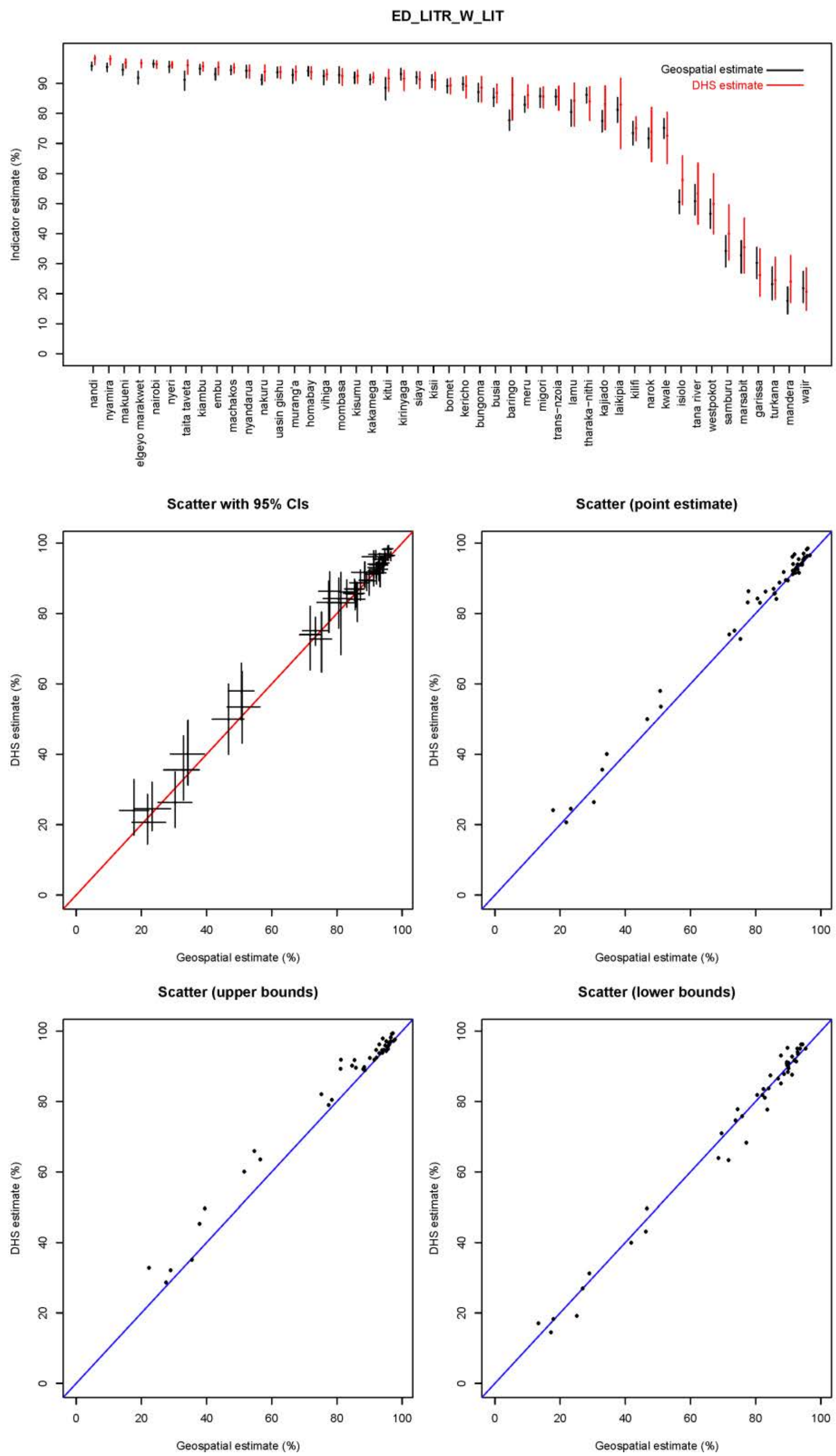


Figure 11G. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is FP_CUSM_W_MOD.

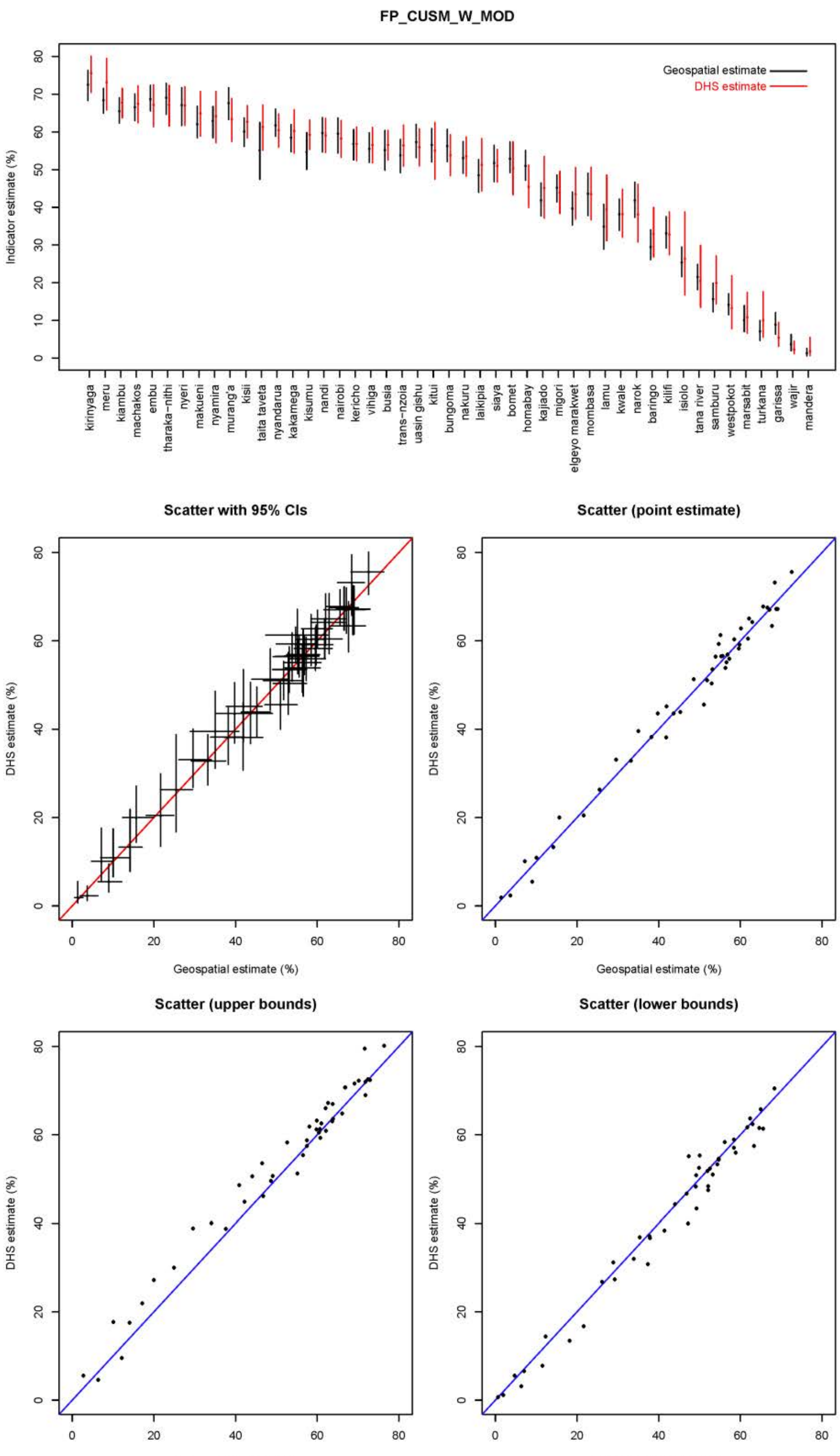


Figure 11H. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is ML_ITNA_P_ACC.

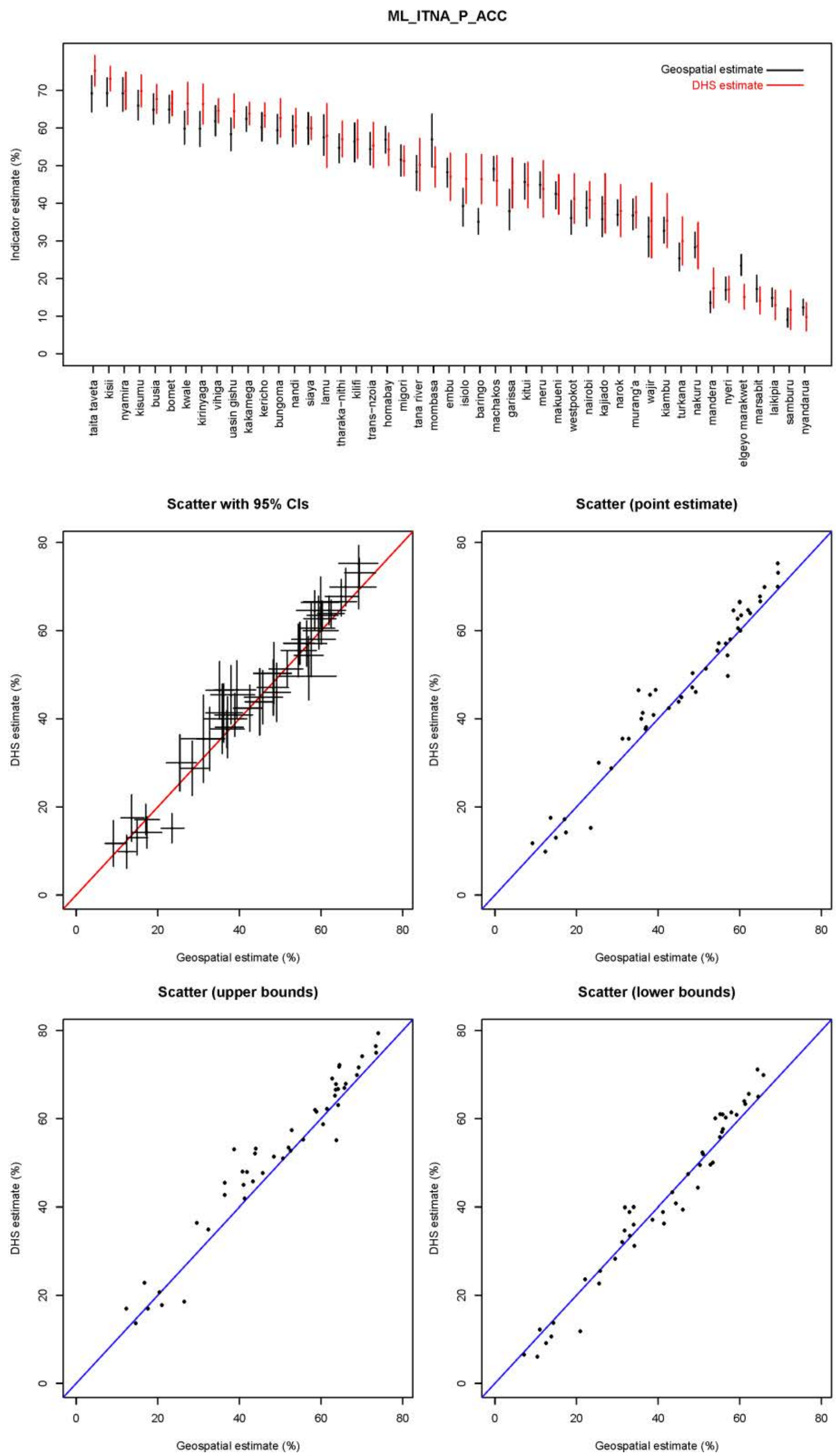


Figure 11l. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is RH_ANCN_W_N4P.

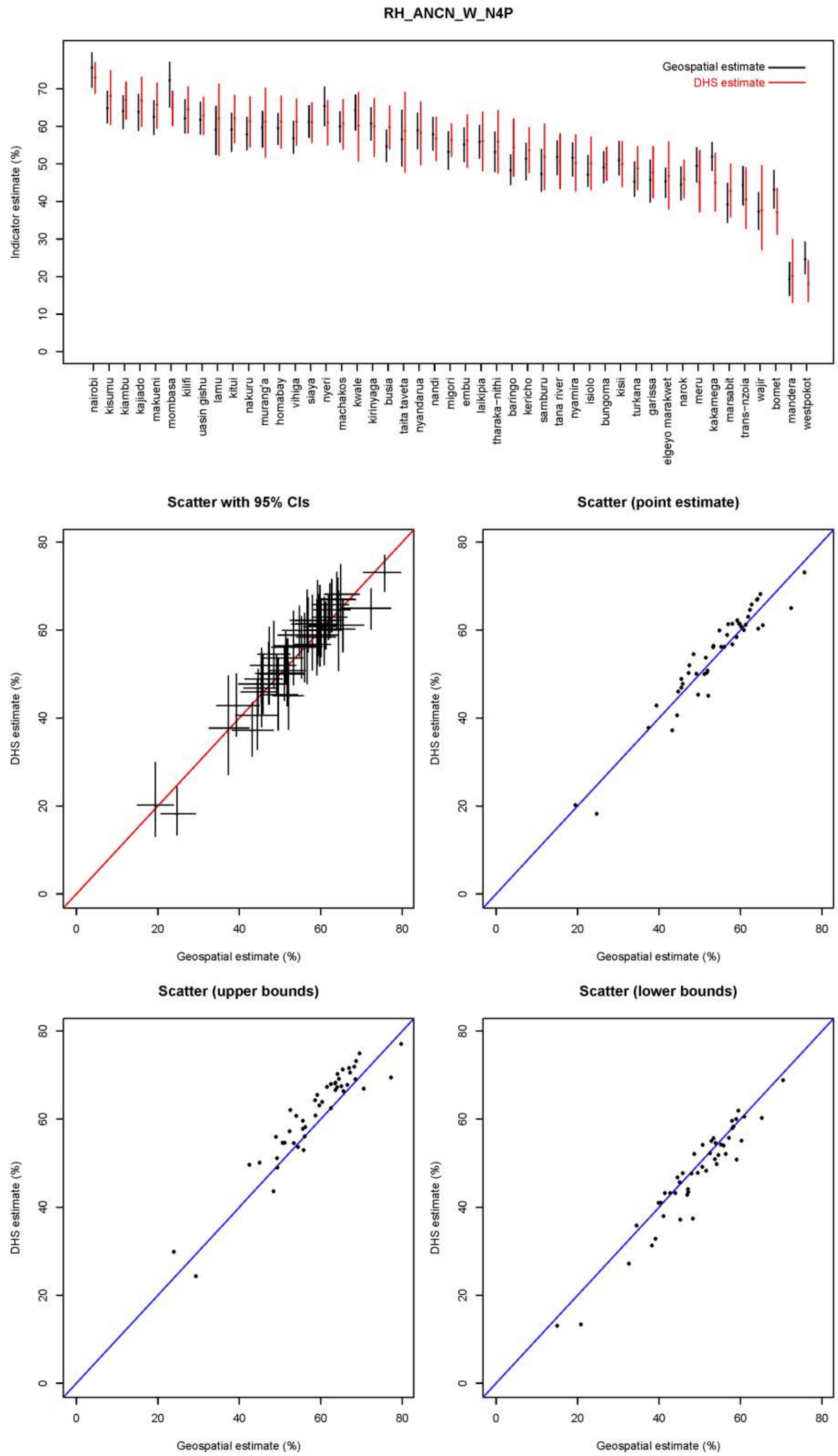


Figure 11J. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is RH_DELP_C_DHF

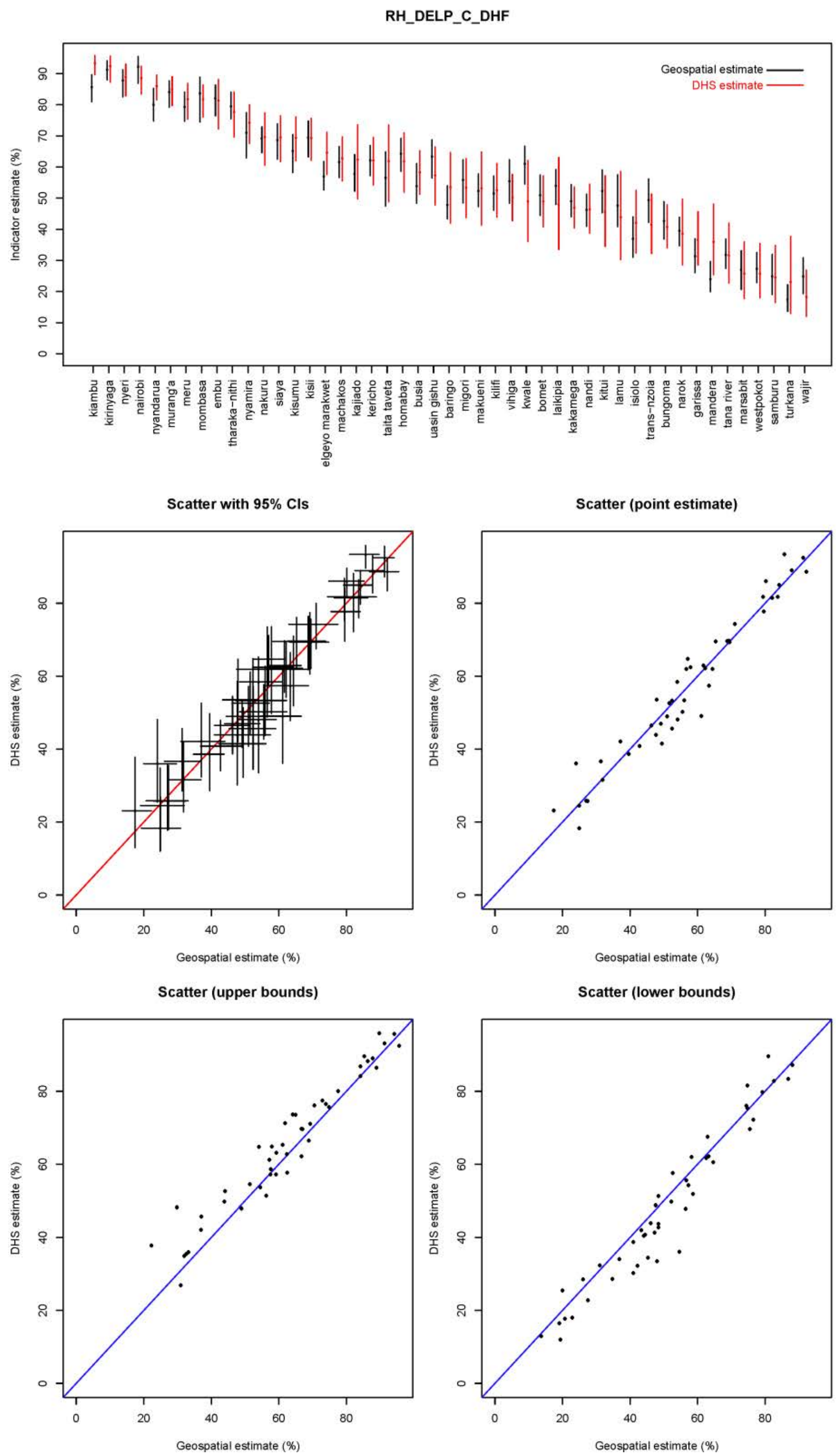


Figure 11K. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is WS_SRCE_P_IMP.

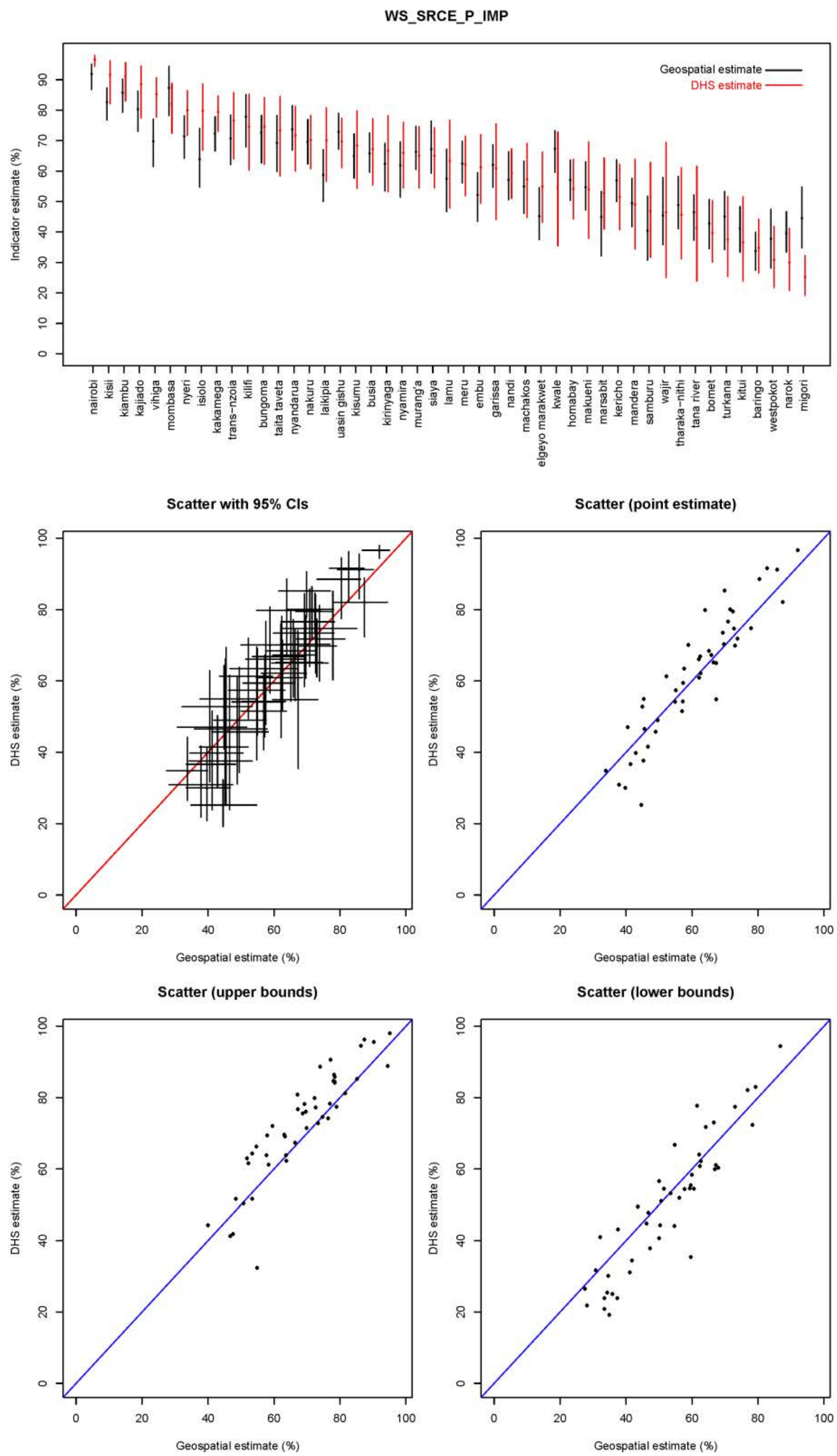
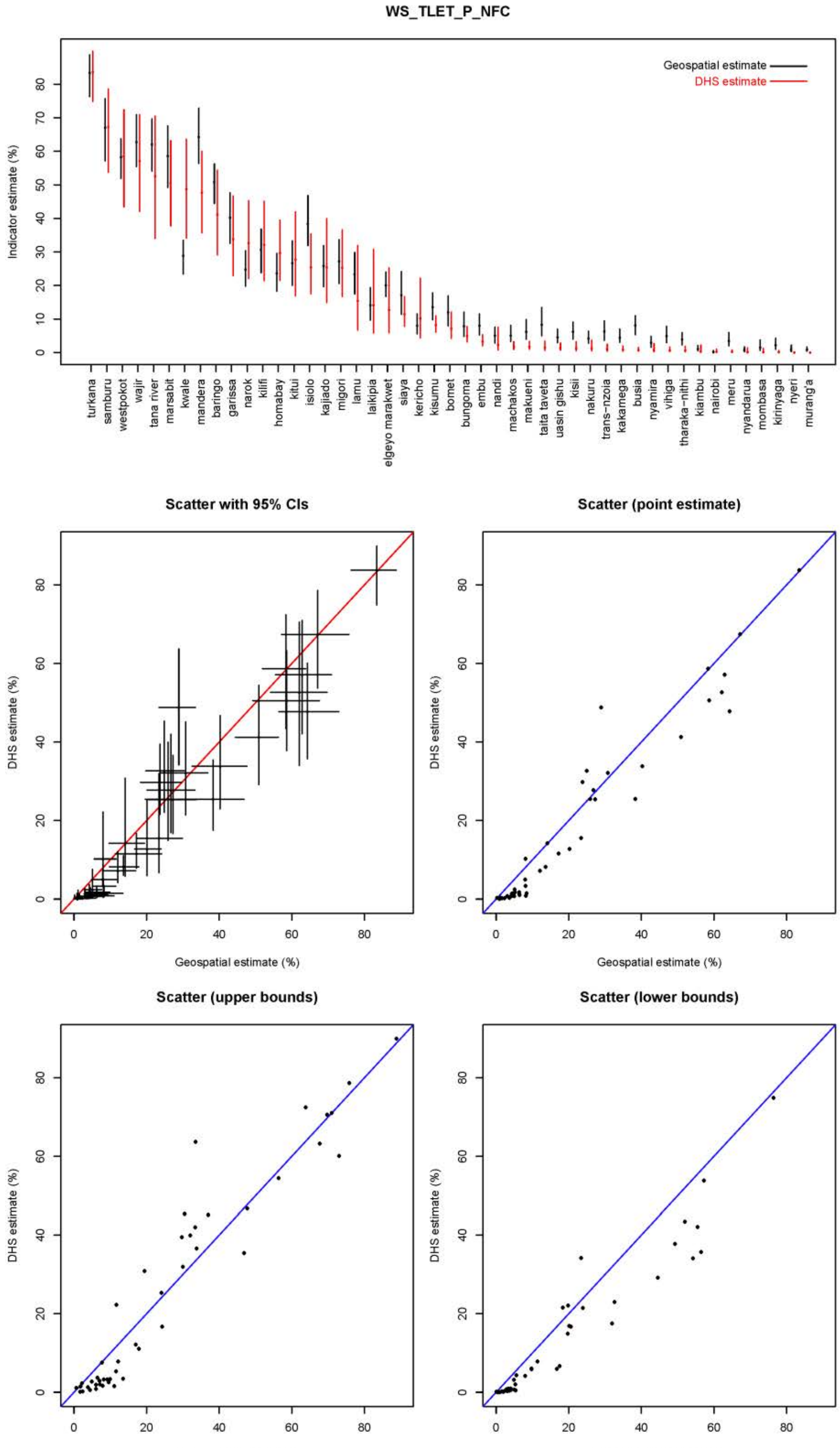


Figure 11L. Comparisons of County-level estimates as derived directly from DHS survey data versus from geostatistical model. Indicator shown here is WS_TLET_P_NFC.



Annex 2. Model-Based Geostatistical Framework for Generating Standardized Modeled Surfaces of DHS Indicators

Full descriptions of the approach are included in SAR11 (Gething et al. 2015) and SAR 14 (Burgert et al. 2016). In brief, a Bayesian model−based geostatistical (MBG) approach (Diggle and Ribeiro 2007; Diggle, Tawn, and Moyeed 1998) was used to generate modeled surfaces for each country-indicator. Building on techniques originally conceived for detailed mapping of malaria prevalence (Gething et al. 2011; Hay et al. 2009), MBG models represent the observed variation in cluster−level survey data using four components.

Sampling error, which can often be large given the small sample sizes in individual clusters, is represented using a standard sampling model, usually the binomial when the indicator in question is a proportion, as is most often the case for DHS indicators.

Some non−sampling variation can often be explained using fixed effects, whereby a multivariate regression relationship is defined linking the indicator variable with a suite of geospatial covariates.

Additional non−sampling errors not explained by the fixed effects are usually spatially auto-correlated, and this is represented using a random effects component. A spatial multi−variate normal distribution known as a Gaussian Process is employed, parameterized by a spatial covariance function.

Any remaining variation not captured by these components is represented using a simple Gaussian noise term equivalent to that employed in a standard non−spatial linear model.

Two types of data are input into the modeled surface process:

1. DHS cluster level observations: Using the publicly available DHS data (individual and household recode files), the cluster level numerator and denominator for the indicator are created. This information is then linked to the cluster level GPS location data.

2. Geospatial covariates: A range of covariate grids are included as possible explanatory covariates. 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. The covariates are standardized to a 5x5km raster grid within a uniform coastline.