

Geo-Spatial Modeling of Access to Water and Sanitation in Nigeria

Luis A. Andres

Samir Bhatt

Basab Dasgupta

Juan A. Echenique

Peter W. Gething

Jonathan Grabinsky Zabludovsky

George Joseph



WORLD BANK GROUP

Water Global Practice

February 2018

Abstract

The paper presents the development and implementation of a geo-spatial model for mapping populations' access to specified types of water and sanitation services in Nigeria. The analysis uses geo-located, population-representative data from the National Water and Sanitation Survey 2015, along with relevant geo-spatial covariates. The

model generates predictions for levels of access to seven indicators of water and sanitation services across Nigeria at a resolution of 1×1 square kilometers. The predictions promise to hone the targeting of policies meant to improve access to basic services in various regions of the country.

This paper is a product of the Water Global Practice. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at <http://econ.worldbank.org>. The authors may be contacted at Landres@worldbank.org.

The Policy Research Working Paper Series disseminates the findings of work in progress to encourage the exchange of ideas about development issues. An objective of the series is to get the findings out quickly, even if the presentations are less than fully polished. The papers carry the names of the authors and should be cited accordingly. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent.

Geo-Spatial Modeling of Access to Water and Sanitation in Nigeria¹

Luis A. Andres, Samir Bhatt, Basab Dasgupta, Juan A. Echenique, Peter W. Gething, Jonathan Grabinsky Zabludovsky, and George Joseph

JEL Classification: C51, C55, J18, Q01, Q25

Key Words: Geo-Spatial Modeling, Water, Sanitation, Sustainable Development Goals, Nigeria.

¹ Luis A. Andres: World Bank Water Global Practice, landres@worldbank.org. Basab Dasgupta: Social Impact, Impact Evaluation Division, bdasgupta@socialimpact.com. Juan A. Echenique: School of Public Policy, University of Maryland, jac@umd.edu. Samir Bhatt: Imperial College London, bhattsamir@gmail.com. Peter W. Gething: Big Data Institute, Nuffield Department of Medicine, University of Oxford, peter.getthing@bdi.ox.ac.uk. Jonathan Grabinsky Zabludovsky: World Bank Water Global Practice, gjoseph@worldbank.org. George Joseph: World Bank Water Global Practice, gjoseph@worldbank.org. The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the view of the World Bank, its executive directors, or the countries they represent. The findings, interpretations, and any remaining errors in this paper are entirely those of the authors.

1. INTRODUCTION

Until now, efforts to measure access to water and sanitation around the world have provided a certain level of aggregation at the subnational level, such as for particular government districts, but rarely do we encounter high-resolution maps for entire countries. Using survey data to map particular indicators is difficult for a number of reasons. First, the actual location of the surveyed establishment is usually unavailable. Second, due to cost constraints, and to ensure representativeness, surveys typically use cluster-based sampling techniques, which makes the distribution of observations uneven across a given area. The absence of reliable, granular, evenly distributed, geo-referenced data makes it difficult to accurately compare water, sanitation and hygiene (WASH) access across a country, or to identify those areas in greatest need of investment.

The poor provision of safe, accessible water and sanitation services in Nigeria has commensurate public health and economic impacts. Evidence from Nigeria has shown that those sectors of the population with the worst water, sanitation, and hygiene conditions are also the ones most at risk of attaining diseases due to inadequate health. A majority share of the Global Burden of Disease (GBD) enteric burden – a common measure for estimating the health burden and risk factors of diseases – estimated for Nigeria is associated with inadequate WASH, and disproportionately borne by poorer children and those in vulnerable geographic areas. Approximately 73 percent of the GBD enteric burden estimated for the country is associated with inadequate WASH.²

A recent, nationwide multi-sector assessment undertaken by the Federal Ministry of Water Resources (FMWR) of the Government of Nigeria with support from the World Bank – the 2015 National Water and Sanitation Survey, NWSS – provides uniquely detailed information on access to WASH in Nigeria, as gathered from a wide-ranging set of surveys: a nationally representative household survey on access to safe water and sanitation which covered 201,842 households, a spatial inventory of 89,721 water points and 5,100 water schemes in the country, and a survey on the provision of WASH in over 50,000 public facilities, including health and educational centers.³

² Andres et al (2017).

³ Please refer to Andres et al. (2017) for more information on the NWSS.

The model presented here makes use of the NWSS household survey, as well as the surveys on water points and water schemes, all of which include geo-locational data.⁴ These data present an unprecedented opportunity to use geo-spatial models to analyze, at a detailed level, the geographical characteristics of access to safe water and sanitation across the country.

In sectors outside WASH, many household and facility surveys now include geo-locational information (e.g., the latitude and longitude of survey clusters, recorded via a Global Positioning System [GPS] device at the time of the survey, or linked to spatial administrative boundary data). Spatial, statistical modeling approaches are being developed by exploiting this locational information to generate mapped surfaces of indicators of interest at increasingly fine spatial scales, and with greater precision than was previously possible. Central to many of these approaches is a body of theory known as model-based geo-statistics (MBG) (Diggle & Ribeiro 2007; Diggle et al. 1998). MBG has been successfully applied to point-located survey data to create a wide range of maps, including, for example, mapping malaria prevalence (Gething et al. 2011, 2012) and poverty (World Bank 2016). The availability of the NWSS 2015 data makes it possible to extend the MBG approach to mapping local populations' access to water and sanitation services, and their proximity to the nearest functioning water source, in Nigeria. The high level of granularity resolved in the mapped outputs can improve our understanding of inequalities in access levels between and within the different regions of the country.

2. DATA

National Water and Sanitation Survey (NWSS) 2015

Data on access to WASH variables come from the 2015 NWSS household survey. The household survey was conducted by the Federal Ministry of Water Resources, which interviewed 201,842 households across 36 states in Nigeria (Figure 1).⁵ The survey asked questions relating to

⁴ All surveyed households and water service points were georeferenced in the surveys to provide latitude and longitude coordinates. Water schemes were also georeferenced using their centroid location, although it should be noted that in many cases these schemes occupy a significant area and so the use of a single central location is a potentially crude approximation of their true spatial extent and coverage.

⁵ See a more detailed description at Andres et al. (2017).

respondents' access to water and sanitation services, and their use of water and sanitation infrastructure. It also included questions on household expenditure, health and hygiene.

From the NWSS household survey , we were able to construct seven access to WASH indicators, informed by the Sustainable Development Goals (SDGs) (WHO/UNICEF 2015). These indicators are: (1) access to improved water, (2) access to basic water, (3) access to improved water on premises, (4) access to piped water on premises, (5) lack of access to fixed-point sanitation (also known as open defecation), (6) access to improved sanitation, and (7) access to sewerage connection, with definitions as follows:

- (1) **Improved water** sources are those which, by the nature of their construction and when properly used, are adequately protected from outside contamination, particularly fecal matter. Such sources include piped water to yards/plots, public taps or standpipes, tube wells or boreholes, protected springs, and rainwater.
- (2) **Basic water** satisfies the requirements of “improved water” while also satisfying the additional requirement that it take less than 30 minutes, round trip, to collect the water in question.
- (3) **Improved water on premises** fulfills the same requirements as basic water, but further implies that the water is available directly on household premises.⁶
- (4) **Piped water on premises** fulfills the same requirements as improved water on premises, but is provided through pipes.
- (5) **Fixed-point sanitation** involves a pit or other containment structure, regardless of the quality of the structure or whether it is hygienically maintained. While it includes both improved and unimproved facilities, it stands in contrast to open defecation, which is defined as not having access to any type of toilet.

⁶The global SDG indicator for water is defined as the “percentage of population using safely managed drinking water services,” and covers those improved drinking water sources that are (1) located on premises, (2) available when needed, and (3) compliant with fecal and priority chemical standards. Unfortunately, at the time the FMWR commissioned data collection for the National Water and Sanitation Survey (NWSS), this SDG indicator had not yet been defined, so we did not include access to safely managed water in the MBG model.

- (6) An *unshared improved sanitation facility*, an indicator of *improved sanitation*, is one that hygienically separates human excreta from human contact and is not shared with any other household.⁷
- (7) *Sewerage* implies that an improved sanitation facility is connected to a sewer system.

Geo-spatial covariates and population data

In addition to the NWSS's outcome data on the indicators of interest, a second category of data used for analysis was a suite of geo-spatial covariates that may be correlated with the indicators of interest, and thus partially explain observed spatial variation, allowing for more accurate predictions across each map. Geo-spatial covariates are gridded spatial data: each grid cell (or pixel) contains the value of a particular property. An initial set of spatial covariates were identified as potentially useful predictors of water and sanitation access levels, based on previous attempts to predict poverty in Nigeria (Gething & Molini 2015). This set of covariates is presented in Figure 2 and consists of (1) a vegetation index, (2) aridity, (3) land-surface temperature, (4) brightness of nighttime lights, and (5) estimated travel time to the nearest functioning water source. The spatial covariates may be described as follows:

- (1) *Vegetation index* (Figure 2a). NASA's Moderate Resolution Imaging Spectroradiometer (MODIS, <http://modis.gsfc.nasa.gov/>) generates high-resolution satellite imagery on various measures of environmental conditions. This includes the Enhanced Vegetation Index (EVI), which measures reflectance in the green and red parts of the visible spectrum to provide a relative measure of the density of photosynthesizing vegetation in each pixel. These data were preprocessed to provide average values for the year 2015 in each 1x1 kilometer (km) pixel.
- (2) *Aridity* (Figure 2b). The Consultative Group for International Agricultural Research (CGIAR) Consortium maintains high-resolution global raster climate data related to evapotranspiration processes and a rainfall deficit for potential vegetative growth. These are based on data from the WorldClim project (Hijman et al. 2005), and ultimately from

⁷ The global SDG indicator for sanitation, "percentage of population using safely managed sanitation services," implies the use of an improved sanitation facility that is not shared with other households, and where excreta are safely disposed on site or transported and treated offsite. Unfortunately, at the time the FMWR commissioned data collection for the National Water and Sanitation Survey (NWSS), this indicator had not yet been defined, so data about excreta disposal or treatment were not collected.

weather station data interpolated using covariates such as altitude (<http://csi.cgiar.org/Aridity/>).

- (3) *Land surface temperature* (Figure 2c). NASA's MODIS also generates high-resolution satellite imagery on land surface temperature (LST).
- (4) *Brightness of nighttime lights* (Figure 2d). This information comes from the Defense Meteorological Satellite Program Operational Linescan System's (DMSP OLS's) annual composite satellite data for nighttime lighting in 2009 (<https://ngdc.noaa.gov/eog/>). These data allow regions to be differentiated by the density of their population and also the degree of the electrification of their dwellings, commercial and industrial premises, and infrastructure.
- (5) *Estimated travel time to nearest functioning water source* (Figure 2e). This covariate was created for the current study by first creating a "friction surface" that estimates the time required to traverse each 1x1 km pixel across Nigeria. This varies according to the type of land cover, topography, and the layout of the road and the wider transport network across the country. The friction surface was then used in a least-cost path algorithm to estimate the likely travel time from the center of each 1x1 km pixel to the nearest functioning improved water source (such as a well, bore hole, or pump). The latitude and longitude, as well as the level of functionality, of every such water point and water scheme in Nigeria was recorded as part of the NWSS 2015.
- (6) A final category of data used in the analysis was a gridded map of estimated population density across Nigeria (Figure 2f) constructed from satellite-derived settlement maps and available census data as part of the AfriPop project (www.afripop.org) (Linard et al. 2012). An alternative population grid, from the Global Rural Urban Mapping Project (GRUMP, <http://sedac.ciesin.columbia.edu/data/set/grump-v1-population-density>) was also investigated. These gridded population surfaces were not used as covariates but were used to calculate population-weighted mean and count estimates for the various modeled indicators.

Defining and implementing a standardized grid format

The geo-spatial data sources described above were obtained in a variety of spatial resolutions and geographic extents. The land-sea templates inevitably varied, so 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 generate a fully standardized suite of input grids on an identically defined geographic template, a processing chain with the following stages was developed. First, each input data source was re-projected, where necessary, using a standardized equiarectangular Plate Carrée projection under the World Geodetic System 1984 coordinate system. Second, where input grids were defined at differing spatial resolutions, they were re-sampled to 1×1 km. Third, grids were either extended or clipped to match a standardized extent. Fourth, a bespoke algorithm was developed that compared each rectified and re-sampled grid to a “master” land-sea template for Nigeria and used a simple interpolation and/or clipping procedure to align new grids to this master template, thus ensuring that all the coastline was perfectly consistent on a pixel-by-pixel basis.

3. METHODOLOGY

Model-based geo-statistics

The predictive approach used in this study to generate fine-scale maps of each water and sanitation indicator across Nigeria was based on a body of statistical theory known as model-based geo-statistics (MBG). In an MBG framework, the observed variation in cluster-level indicator values is explained by one of the following four components:

- (1) A *sampling error*, which can often be large given the small sample sizes of individual clusters, is represented using a standard sampling model (e.g., a binomial model where cluster-level data consist of a selection of “poor” households from the total number sampled).
- (2) Some non-sampling variation can often be explained using fixed effects – whereby a multivariate regression relationship is defined by linking the dependent poverty variable with a suite of geo-spatial covariates.
- (3) An additional non-sampling error not explained by the fixed effects is usually spatially auto-correlated, and this is represented using a random effect component. A spatial multivariate normal distribution known as a Gaussian Process is employed, parameterized by a spatial covariance function.

(4) Finally, any remaining variation not captured by these components is represented using a simple Gaussian noise term, equivalent to that employed in a standard spatial linear model. The full model output is, for every pixel on the mapped surface, a posterior distribution for the predicted indicator, representing a complete model of the uncertainty around the estimated value. These can be summarized using a point estimate (such as the posterior mean) to generate a mapped surface of the indicator value. This methodology is able to present smaller points of estimation (in the spatial dimension) than are other methodologies such as small area estimation (Blankespoor & van der Weide 2017).

Formal description of the model structure

MBG models are a class of generalized linear mixed models, with an approximation of a multivariate normal random field (i.e., a Gaussian Process) used as a spatially auto-correlated random effect term. Each indicator (the proportion of individuals with access to the specified water/sanitation services) $Y(x_i)$ at each location in Nigeria x_i for the year 2015 was modeled as a transformation $g(\cdot)$ of a spatially structured field superimposed with additional random variation $g(\cdot)$. The count of individuals with access 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 component $\epsilon(x_i)$ was represented as Gaussian with a zero mean and variance V . Both the inference and prediction stages were coded using the Integrated Nested Laplace Approximation (INLA) framework, primarily in the R programming language.

The mean component, μ , was modeled as a linear function of the n geo-spatial covariates, $\mu = \beta x$, where $X = (1, X_1(x), \dots, X_n(x))'$ was a vector consisting of a constant and the covariates indexed by spatial location x , and $\beta = (\beta_0, \beta_1, \dots, \beta_n)$ was a corresponding vector of the regression coefficients. Each covariate was converted to z -scores before analysis. Covariance between spatial locations was modeled using a Matern covariance function:

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

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

Incorporation of covariates

In a standard non-spatial generalized linear model (GLM) regression approach, it is necessary to undertake a formal covariate selection procedure to maximize the ultimate predictive accuracy of the model. Including too few informative covariates means that exploratory power is lost, but the inclusion of too many may result in the high-dimensional multivariate model overfitting the data, explaining noise rather than signal and, ultimately, reducing predictive accuracy. Because full geo-statistical models are extremely time-consuming to fit, a common practice has been to use 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 often used, whereby a covariate set is built up by progressively adding new candidate covariates to a model (forward selection) or subtracting them from an initial inclusive set (backward selection), and deciding to keep or discard each new covariate based on its impact on the model fit. These techniques are, however, known to be sensitive to the order in which variables are added or removed, and therefore risk generating arbitrary final selections.

In this study, a more novel approach has been implemented: the use of “regularization” embedded within the geo-statistical model itself. In intuitive terms, this allows a large suite of candidate covariates to be entered into the main model while achieving two things. First, it allows the model to sacrifice a small amount of bias for a large reduction in variance (in a trade-off between bias and variance), greatly improving out-of-sample predictive capacity. Second, the regularizer shrinks the coefficients of the covariates, which means the effects of collinearity are minimized, making the model more stable and robust. In formal terms, a Gaussian process anterior was imposed on the likelihood, allowing regularization of the posterior mean:

$$p(f_x|y) = \frac{p(y|f_x)p(f_x)}{p(y)} = \frac{N(y; f_x, \sigma^2 I)N(f_x; \mu, C)}{N(y; \mu, C + \sigma^2 I)}$$

$$-2\log p(f|y) = (y - f_x)^T \sigma^{-2} I (y - f_x) + (f_x - \mu)^T C (f_x - \mu) + \text{constant}$$

$$-2\log p(f|y) = \sigma^{-2} \|y - f_x\|_I^2 + \|f_x - \mu\|_C^2 + \text{constant}$$

Here $N(\cdot)$ is the Gaussian probability distribution function; f_x is the Gaussian process function; y is the response; μ, C are the mean and covariance functions, as defined earlier; and $\sigma^2 I$ is the noise or error. The regularization is not just the l_2 distance in the conventional ridge regression but the Mahalanobis distance, which accounts for the elliptical skew due to the covariance function, thereby including all correlated effects into the regularizer. In addition to the conceptual benefits afforded by the Gaussian process prior, the possible inclusion of *a priori* non-linear transformations on the fixed effects was explored. However, these non-linear transformations did not lead to significant improvements over the non-transformed parsimonious model, and so the latter was retained. Model complexity was measured using the Deviance Information Criteria.

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 1×1 km spatial grid across Nigeria and of the unobserved parameters of the mean, covariance function, and Gaussian random noise component. At each location, the posterior distribution was summarized using the posterior mean as a point estimate, and maps were generated of each of these metrics in ArcGIS 10.4.

Aggregation at the level of individual states and local government areas (access rate and count)

The MBG models generate predicted maps of each indicator at a 1×1 km resolution. While these provide the most fine-grained picture of variation in water and sanitation access across the country, it is also useful to summarize these patterns at higher levels of aggregation corresponding to the administrative unit levels at which program planning, implementation, and decision-making are

carried out. For each indicator, therefore, various aggregate versions were calculated at both the level of the state (1st subnational unit) and local government area (LGA, 2nd subnational unit), as follows:

- (1) *Mean indicator rates*. These are calculated as population-weighted means of the indicator predictions across all pixels within each administrative unit, and provide the best estimate of the percentage of the population within each unit that meets the criterion of each indicator (e.g., the percentage of people with access to basic water in state x).
- (2) *Indicator rate quintiles*. Mapping the mean indicator rates allows for a comparison of the absolute level of access across administrative units. Also of interest is the *relative* level of access, and this is best visualized by identifying the quintile within which each administrative unit lies relative to others across the country.
- (3) *Indicator count*. This is the sum of the population in each administrative unit that meets the criterion for the indicator. Since this metric is primarily used to help target underserved populations, a count was calculated for that fraction of the population *without* access to water/sanitation services (e.g., the count of people that *do not have* access to basic water in state x).

4. RESULTS

Model coefficients

Table 1 shows fitted coefficients for each of the fixed effects (covariates) used in the model for each water and sanitation indicator. Since these are Bayesian models, each parameter is estimated as a full posterior distribution, and is summarized here via the 50th (median), 2.5th, and 97.5th percentiles. The magnitude, direction, and significance of fitted coefficients varied considerably across the different indicators. In some cases, the observed relationships matched prior expectations: for example, that access to basic and improved water was inversely correlated to an increase in travel time to the nearest water point or scheme, or that areas that were more lit up at night (thus more urban) were associated with higher access to sewerage connections and piped water on premises, and lower rates of open defecation. Others were less intuitive: for example, that improved sanitation rates were higher in areas that were less bright at night. It should be noted

that, although many covariates contributed in a statistically significant way to the final model fits, their interpretation is not as straightforward as in a non-spatial model, because much of the variation in observed indicator values is accounted for via the random effect component.

Table 1 Parameter estimates for fixed effects (covariates)

	Percentile	EVI	Aridity	LST	NTL	Time to waterpoint
Basic water	2.5 th	-0.699	-2.290	-4.184	-0.064	-0.033
	50 th	-0.041	-0.765	-2.066	-0.020	-0.028
	97.5 th	0.615	0.758	0.051	0.024	-0.022
Improved water	2.5 th	-1.597	-2.226	-4.409	-0.084	-0.050
	50 th	-0.869	-0.506	-2.061	-0.036	-0.044
	97.5 th	-0.143	1.209	0.282	0.011	-0.039
Improved water on premises	2.5 th	-0.353	-2.181	-4.438	-0.104	-0.005
	50 th	0.248	-0.782	-2.497	-0.064	0.000
	97.5 th	0.849	0.611	-0.559	-0.024	0.005
Piped water on premises	2.5 th	1.251	-1.699	-2.541	0.012	0.014
	50 th	1.723	-0.618	-1.055	0.042	0.017
	97.5 th	2.197	0.456	0.432	0.072	0.021
Open defecation	2.5 th	4.083	-3.472	-1.945	0.151	0.014
	50 th	4.728	-1.899	0.121	0.191	0.018
	97.5 th	5.373	-0.324	2.187	0.231	0.023
Improved sanitation	2.5 th	-0.628	-1.857	-3.760	-0.096	-0.005
	50 th	0.012	-0.343	-1.692	-0.054	0.000
	97.5 th	0.652	1.168	0.377	-0.012	0.005
Sewerage connection	2.5 th	1.908	-3.598	-5.114	0.013	0.012
	50 th	2.320	-2.650	-3.803	0.039	0.015
	97.5 th	2.731	-1.697	-2.489	0.064	0.018

Note: EVI, enhanced vegetation index; LST, land surface temperature; NTL, brightness of nighttime lights. In a Bayesian model, each coefficient is fitted as a probability distribution function, and this is summarized here by the median and 95% credible interval range. Coefficients statistically different from zero (“significant” with 95% confidence) are highlighted in gray.

Model validation

The predictive performance of the model for each indicator is assessed via out-of-sample cross-validation. A fourfold hold-out procedure was implemented whereby 25% of the data points were randomly withdrawn from the data set, 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 with their observed values. This was repeated four times without replacement such 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), accuracy (mean square error), and linear association (correlation) between observed and predicted values.

Table 2 displays validation statistics from the fourfold out-of-sample validation procedure implemented for each predicted variable. The correlation between observed and predicted values was generally very high, exceeding 0.8 (on a scale from zero to one) for most indicators. The two exceptions were piped water on premises and sewerage connection, and here the lower correlations can be attributed to the almost universally low observed values of these indicators – meaning correlations were being assessed within a very small range. Mean absolute errors, which measure the overall precision of the model (and are expressed here on the same scale as the variables themselves – i.e., a proportion between zero and one) again suggested good model performance: the average difference between observed and predicted values at each location was between 0.1 and 0.2. The most precise predictions were for piped water on premises and sewerage connection – again reflecting the lack of variability in the observed data. Mean square errors, which capture overall model performance (both bias and variance), were also small, exceeding 0.05 for only one variable – improved water.

Table 2 Validation statistics summarizing performance of geo-statistical models predicting each water and sanitation variable

Variable	Correlation	Mean absolute error	Mean squared error
Basic water	0.816	0.172	0.047
Improved water	0.830	0.185	0.054
Improved water on premises	0.808	0.142	0.035
Piped water on premises	0.516	0.085	0.014
Improved sanitation	0.815	0.150	0.039
Open defecation	0.865	0.152	0.043
Sewerage connection	0.241	0.076	0.009

Model uncertainty

While the out-of-sample validation procedure provides an external check on the model’s predictive performance and fit, the framework also provides an internal, model-based estimate of the uncertainty associated with the prediction in every pixel. It reveals which parts of each map are more or less certain, as driven by local heterogeneities in the indicator data and the density of data

points. Figure 3 presents uncertainty levels for the water indicators. The estimation results for the indicators of access to improved water, basic water, and improved water on premises show high levels of confidence in densely populated areas. In areas where population numbers are low the precision of the estimates is low. This is favorable from the policy perspective, since certainty is most important in policy decisions that affect the greatest number of people. In the case of piped water on premises, the estimation results have a high level of certainty across a large proportion of the territory. In Figure 4, the results for sanitation indicators are similar to those for water. In the case of indicators with relatively widespread coverage, such as open defecation and improved sanitation, the results again have low levels of uncertainty in areas with high densities of population. For the access to sewerage indicator, at only 5.6 percent, on average, across the nation, a high level of confidence is seen nationwide.

Geo-spatial modeling of basic indicators

In Figures 5–11, the results of the geo-statistical modeling exercise are presented for the seven water and sanitation indicators listed earlier. Each of these figures is divided into three different maps: (1) a detailed pixel-level map shows the predicted percentage of the population, in each 1x1 km pixel, with access to the indicator in question; (2) equivalent percentage estimates are aggregated at the state level; and (3) a population count of those with access to the indicator is defined for each state.

Figure 5 maps the share of population using improved water. The 1x1 km pixel maps reveal pronounced spatial heterogeneity, and across relatively short distances. This is partly due to urban-rural gradients: urban areas tend to have high rates of access to improved water, and rates drop off rapidly outside city limits. At the state level, rates span the range from just 23% (in Bayelsa) to 89% (in Jigawa). The largest concentrations of population without access to improved water are found in Kano (6.0 million), Kaduna (4.5 million), and Benue (3.8 million). Figures 6 and 7 map the share of population using basic water and improved water on premises, respectively. Unsurprisingly, estimated rates are lower for both indicators than for improved water, given their more stringent requirements. Both maps have a similar urban-rural pattern characterized by higher rates of access within and around the major urban centers (especially Lagos and Imo to the north

and Kano to the south). The degree to which these higher urban rates extend past city limits and into surrounding rural areas is far smaller for basic water and improved water on premises than for improved water, leading to a more focal, concentrated urban effect.

At the state level, Enugu has the lowest rates of access to both basic water and improved water on premises (7.5% and 6%, respectively), while Lagos has the highest (75% and 60%, respectively). Interestingly, despite having the highest rates of access, the large urban states also have the largest number of people *without* access. The two largest populations without basic water are in Kano (8 million) and Kaduna (5 million); the largest without improved water on premises are in Kano (9 million) and Lagos (5.5 million). Figures 8 and 9 map the populations with piped water on premises and with a sewerage connection, respectively. Very few Nigerians have access to either: the maps show almost uniform, very low rates nationwide other than in a handful of pockets with some access. Even in the states with the highest access rates (Abuja and Lagos), only 16% and 12% of the population have piped water and sewerage connections, respectively. Only seven states have rates of 10% or more for piped water (Abuja, Plateau, Taraba, Delta, Yobe, Nassarawa, and Jigawa) and just four states have rates of 10% or more for sewerage connections (Lagos, Abuja, Nassarawa, and Taraba).

Figure 10 maps the share of the population using an improved sanitation facility. Here, the spatial pattern is rather different from the others; while there are predominately low rates throughout much of the country, the pixel-level map shows areas of much higher access across the states of Kaduna and Niger and parts of Kano and Jigawa. Interestingly, these well-served areas are not well identified in the state-level aggregate maps, highlighting the importance of looking at variations at a local-level resolution. Rates vary at the state level from 7% in Bayelsa to 57% in Kaduna: the largest populations without access are found in Lagos (12 million) and Kano (8 million). When we compare these results with Figure 9, which shows the predicted level of access to sewerage, we observe that the main difference is in access to improved sanitation. In the case of sewerage, the level of access is very low across all the regions of Nigeria.

Finally, Figure 11 maps the share of population practicing open defecation. This is the indicator that displays perhaps the most polarization across the country: around one-third of states display

very high rates of open defecation, especially in the central and southern areas, excluding the coastal regions. The remainder of the country to the north displays very low rates. Accordingly, the state with the highest rates is Kwara, where 63% of the population practices open defecation, while the practice is least prevalent in Kano, at just 2%.

5. CONCLUSION

In order to design targeted policies, access to geographically specific information is crucial. However, this information is usually derived from representative surveys, whose sampling techniques are meant to save on costs while ensuring the representativeness of the population, but only permit a limited degree of desegregation, so the inferences are not extended to outliers. Geo-spatial models can help address these limitations by generating predictions for areas where information is lacking. In this paper, we implement a model-based geostatistical (MBG) prediction of access to specified water and sanitation services in Nigeria. Using information from households and water points and water schemes gathered as part of the National Water and Sanitation Survey 2015, as well as an array of geo-spatial covariates, we generate layers of information for seven key indicators of access to WASH, at a spatial resolution of 1x1 km.

The availability of these spatially detailed estimates provides a new trove of important information to support the targeting of programs advancing water and sanitation access in Nigeria, and offers more detailed, granular estimates, for tracking progress toward the SDGs.

References

- Andres, L., Duret, M., Mantovani, P., Molini, V. & Ort, R. 2017 *A Wake Up Call: Nigeria Water Supply, Sanitation, and Hygiene Poverty Diagnostic*. World Bank, Washington, DC, USA.
- Blankespoor, B. & van der Weide, R. 2017 *Mapping Access to Water and Sanitation Using Small Area Estimation Methods: With Applications to Bangladesh and Nigeria*. Manuscript.
- Diggle, P. J. & Ribeiro, P. J. 2007 *Model-based Geostatistics*. In: *Springer Series in Statistics*, P. Bickel, P. Diggle, S. Fienberg, U. Gather, I. Olkin & S. Zeger (eds). Springer, New York, USA.
- Diggle, P. J., Tawn, J. A. & Moyeed, R. A. 1998 Model-based geostatistics. *Applied Statistics* 47 (3), 299–326.
- Gething, P. W. & Molini, V. 2015 *Developing an Updated Poverty Map for Nigeria*. Report prepared for the World Bank, Washington, DC, USA.
- Gething, P. W., Patil, A., Smith, D. L., Guerra, C. A., Elyazar, I. R. F., & Johnston, G. 2011 A new world malaria map: Plasmodium falciparum endemicity in 2010. *Malar J.* 10, 378.
- Gething, P. W., Elyazar, I. R. F., Moyes, C. M., Smith, D. L., Battle, K. E. & Guerra, C. A. 2012 A long neglected world malaria map: Plasmodium vivax endemicity in 2010. *PLoS Negl Trop Dis.* 6 (9), e1814.
- Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., & Jarvis, A. 2005 Very high resolution interpolated climate surfaces for global land areas. *Int. J. Climatol.* 25 (15), 1965–1978.
- Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., Tatem, A. J. 2012 Population distribution, settlement patterns and accessibility across Africa in 2010. *PLoS ONE* 7 (2), e31743. doi:10.1371/journal.pone.0031743.
- NASA EOSDIS Land Processes DAAC, USGS Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota (<https://lpdaac.usgs.gov>), accessed [April 15, 2017], at [http:// dx.doi.org/\[DOI\]](http://dx.doi.org/[DOI]).
- Trabucco, A. & Zomer, R. J. 2009 Global aridity index (global-aridity) and global potential evapotranspiration (Global-PET) geospatial database. CGIAR Consortium for Spatial Information. Published online, available from the CGIAR-CSI GeoPortal at: <http://www.csi.cgiar.org/>.
- WHO/UNICEF. 2015 Update and MDG Assessment. WHO Press, World Health Organization, Geneva, Switzerland, 90. <http://doi.org/10.1007/s13398-014-0173-7.2>
- World Bank. 2016 *Poverty Reduction in Nigeria in the Last Decade*. <https://openknowledge.worldbank.org/handle/10986/25825>.

Figure 1 Map showing geo-positioned data from the 2015 National Water and Sanitation Survey on surveyed households (left) and water service points and schemes (right).

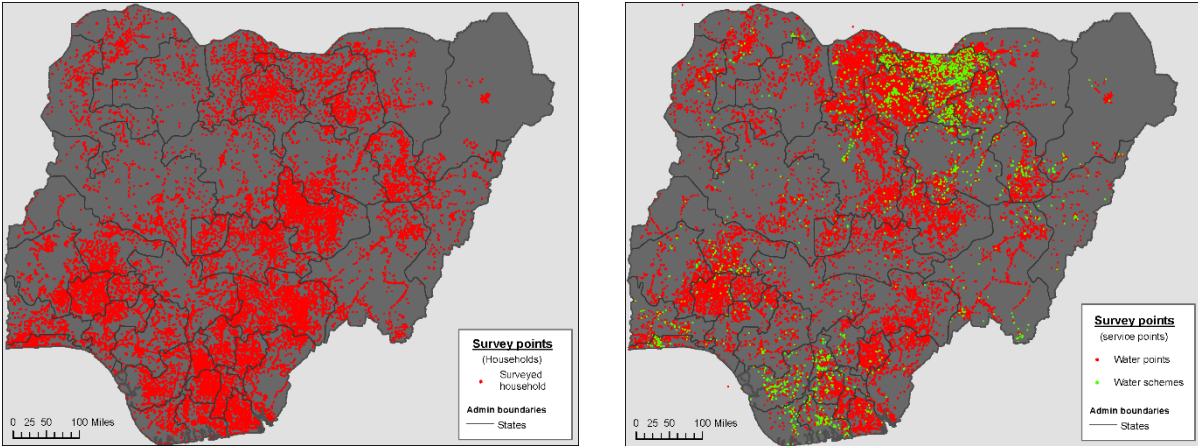
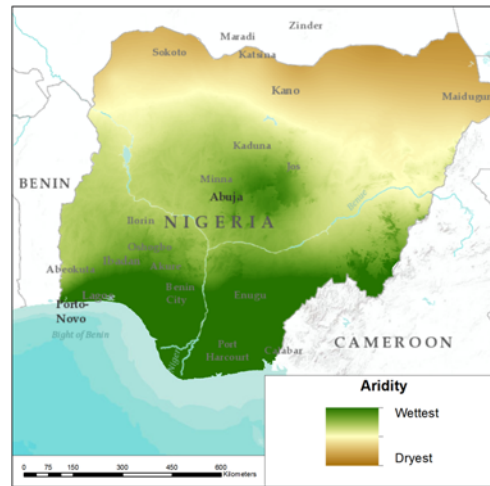


Figure 2 Geo-spatial covariates and ancillary data included in the analysis

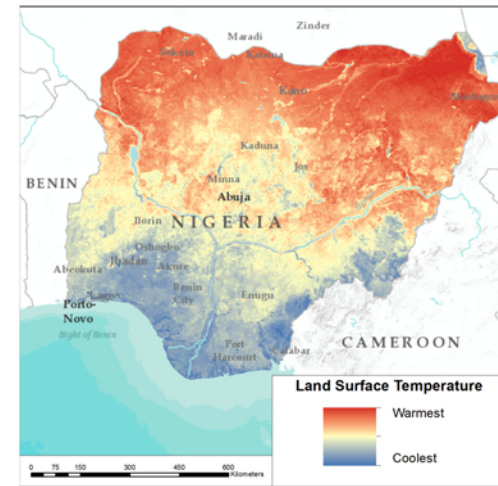
(a) mean enhanced vegetation index imagery derived from NASA's MODIS



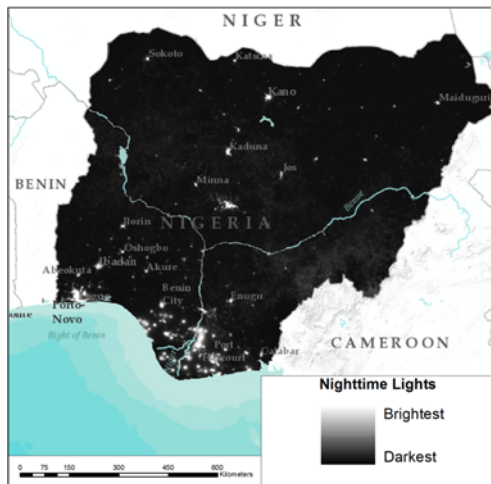
(b) aridity, derived from weather station data and maintained by the CGIAR Consortium



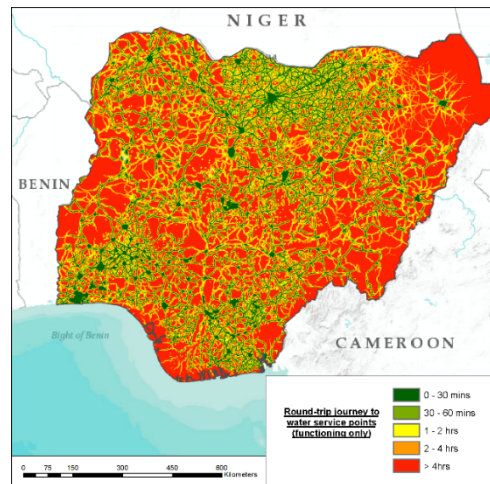
(c) mean land surface temperature from NASA's MODIS



(d) imagery of nighttime lights in Nigeria in 2009 maintained by NOAA



(e) estimated travel time to nearest functioning water service point, as identified in the NWSS 2015



(f) population density layer for Nigeria in 2011 maintained by the AfriPop project

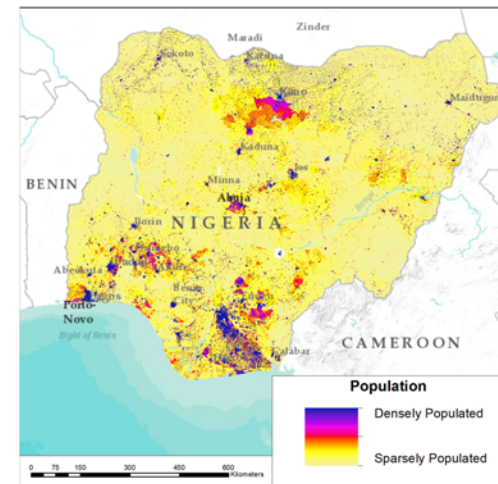
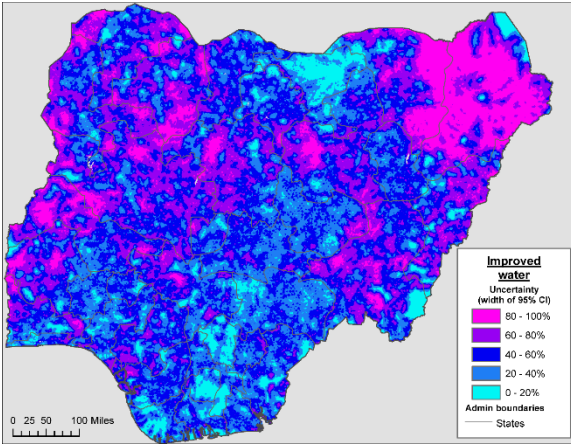
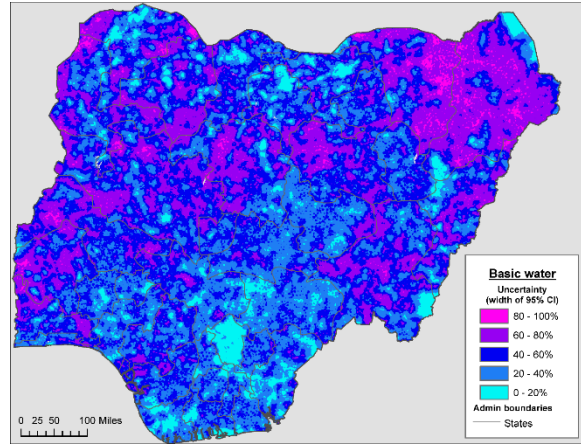


Figure 3 Map showing uncertainty associated with modeled 1×1 km pixel level predictions of the percentage of population with access to four water service indicators. Uncertainty is quantified using the width of the posterior predictive distribution for each pixel (measured on the same scale as the indicator itself: a percentage between 0 and 100%). This is the range of values within which there is a 95% probability that the true indicator value lies, thus wide intervals are more uncertain and narrow intervals less uncertain.

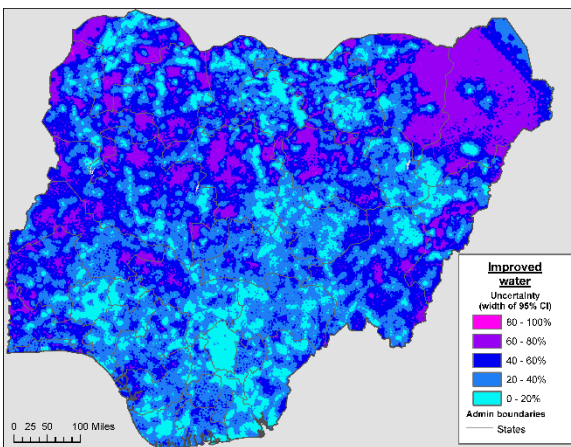
(a) Improved water



(b) Basic water



(c) Improved water on premises



(d) Piped water on premises

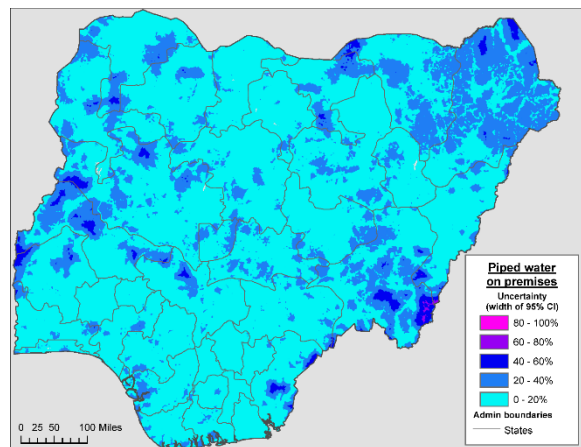
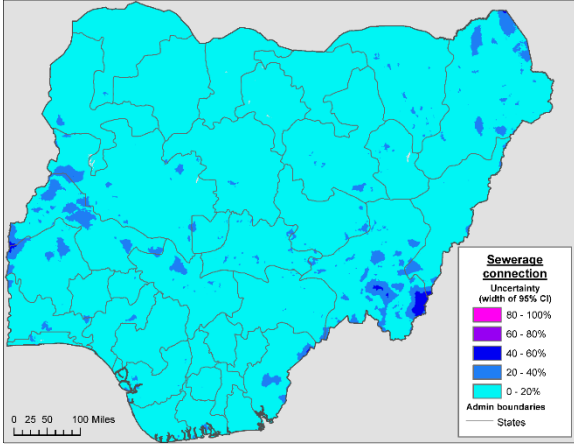
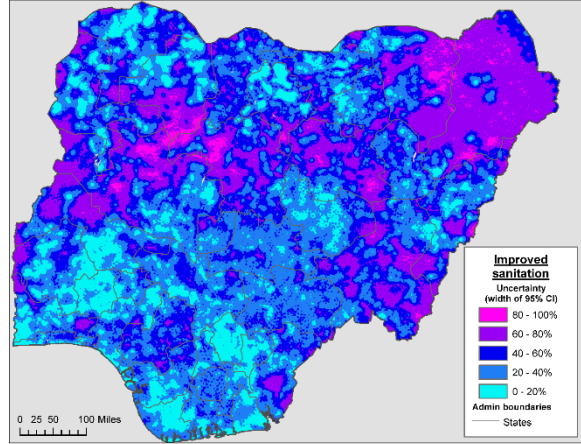


Figure 4 Map showing uncertainty associated with modeled 1×1 km pixel level predictions of the percentage of population with different sanitation access indicators. Uncertainty is quantified using the width of the posterior predictive distribution for each pixel (measured on the same scale as the indicator itself: a percentage between 0 and 100%). This is the range of values within which there is a 95% probability that the true indicator value lies, thus wide intervals are more uncertain and narrow intervals less uncertain.

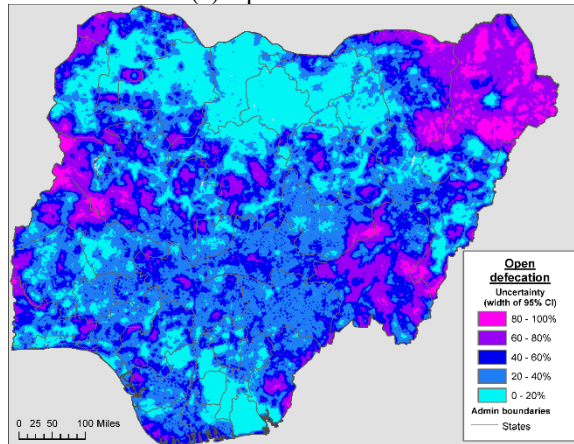
(a) Sewerage connection



(b) Improved sanitation



(c) Open defecation



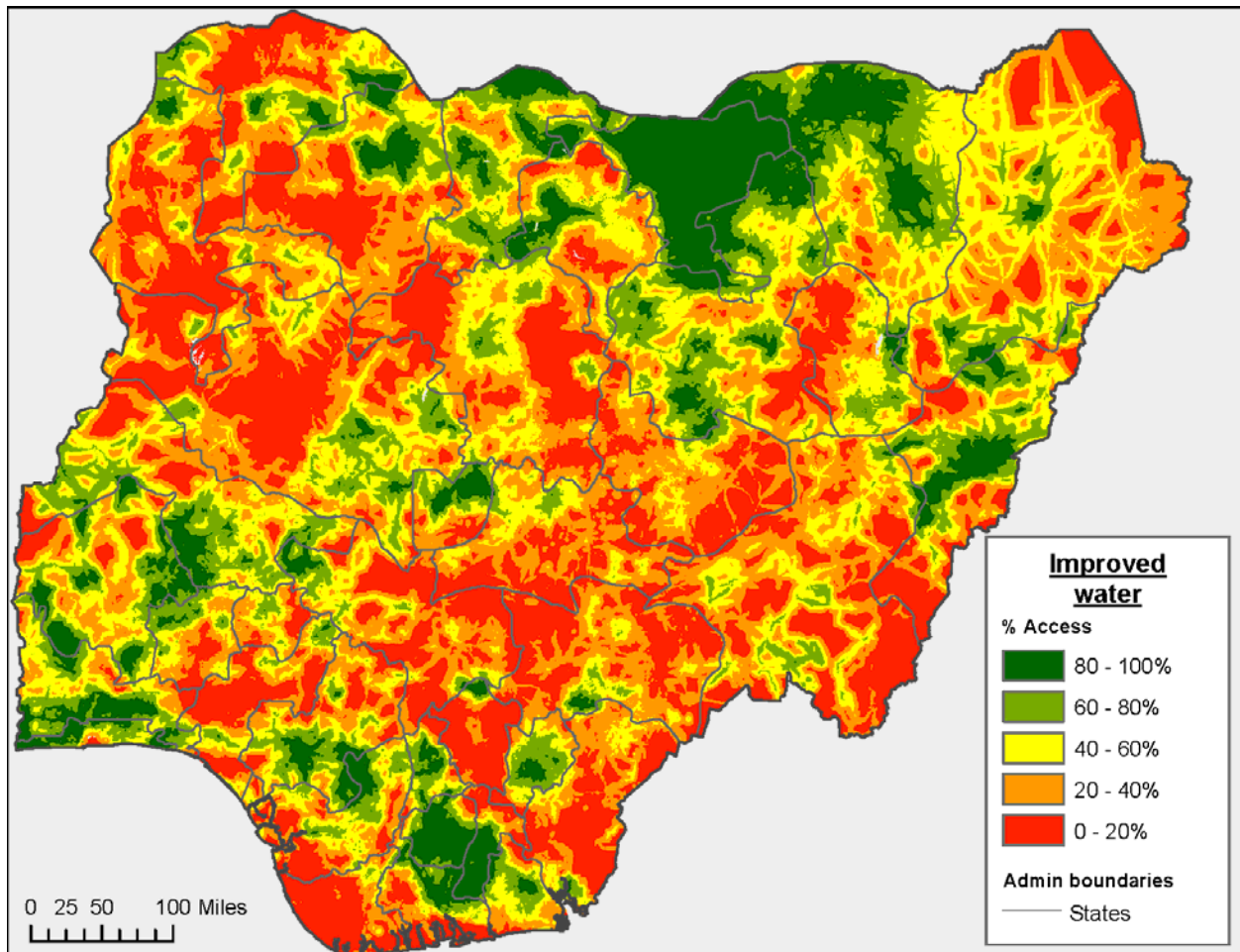
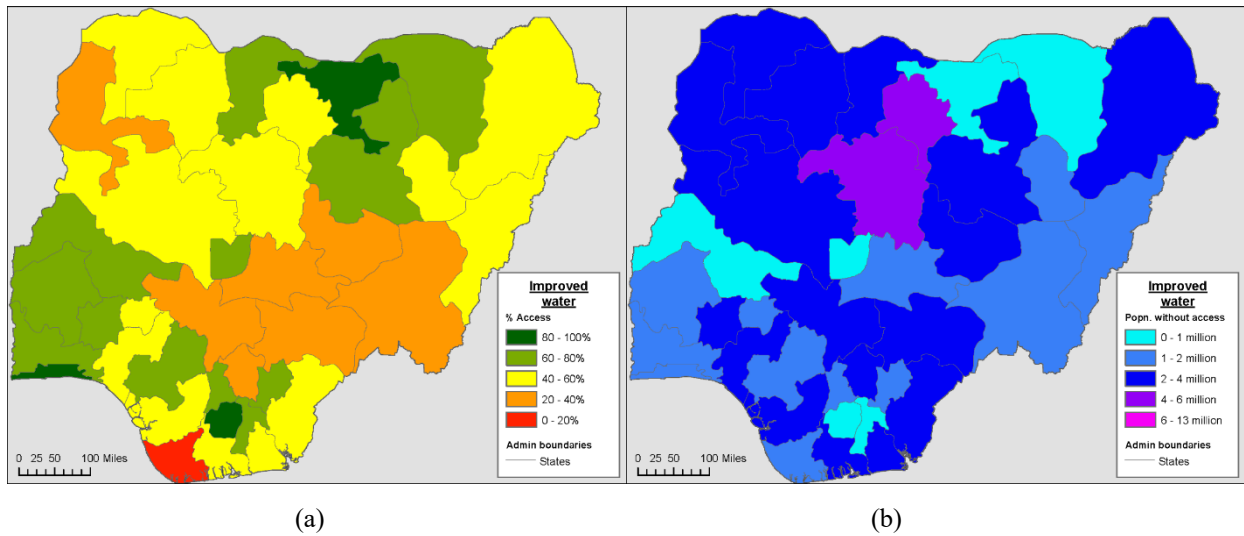


Figure 5 (main) Map showing modeled 1×1 km pixel level predictions of the percentage of population using improved water. Also shown are state-level estimates of (a) the percentage of people with improved water and (b) the number of people without improved water.

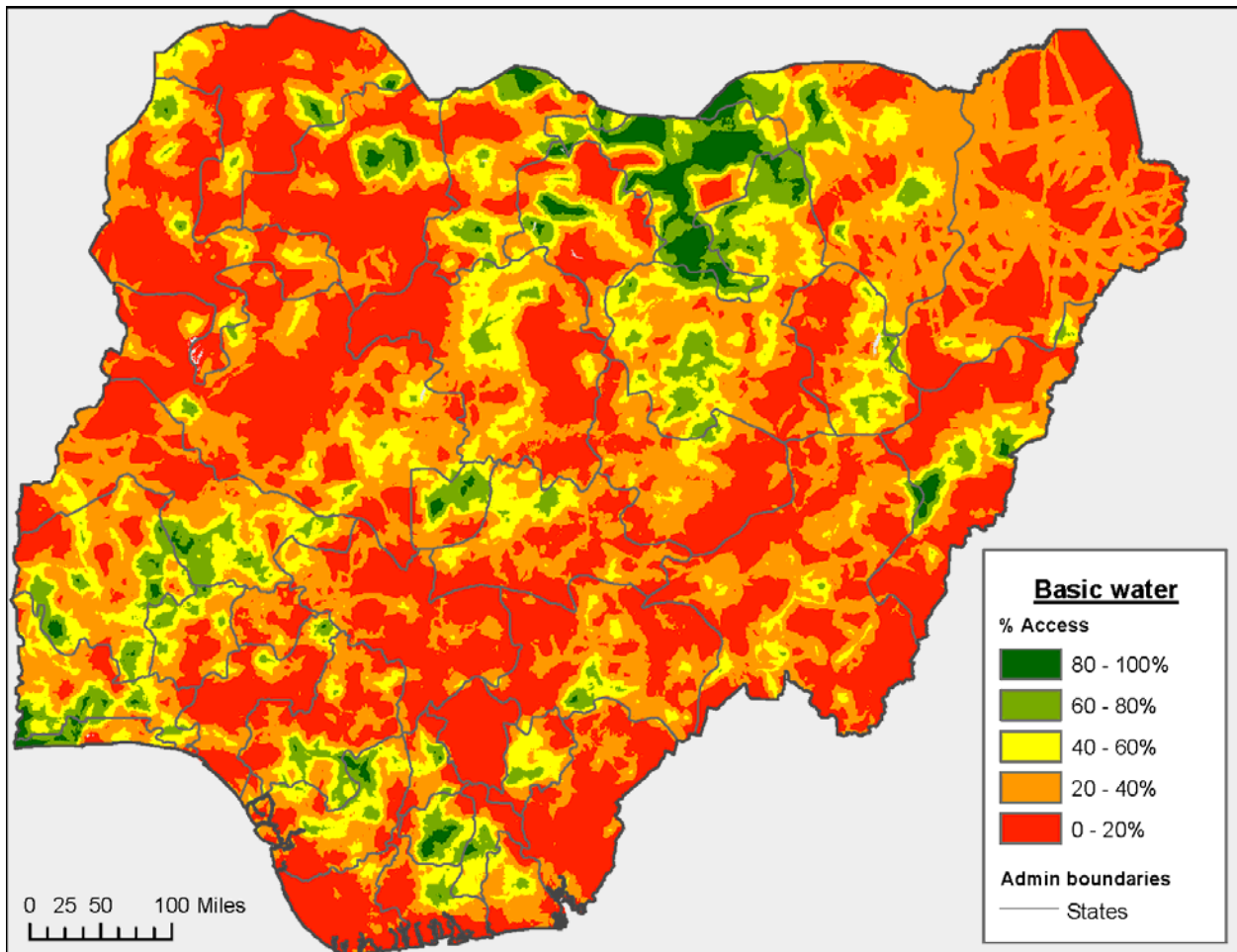
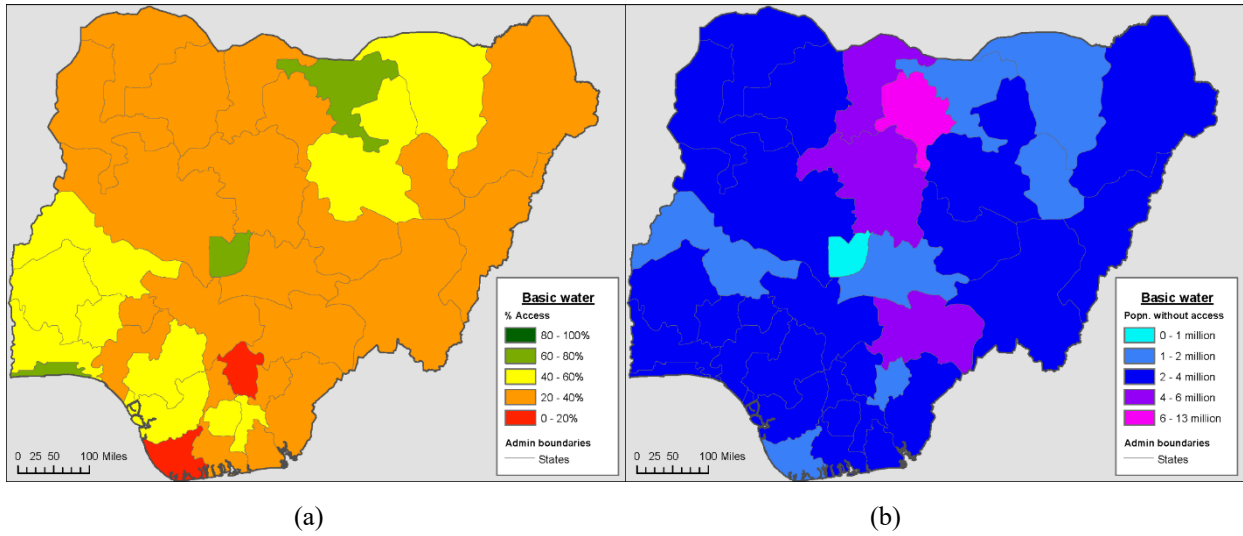


Figure 6 (main) Map showing modeled 1×1 km pixel level predictions of the percentage of population using basic water. Also shown are state-level estimates of (a) the percentage of people with basic water and (b) the number of people without basic water.

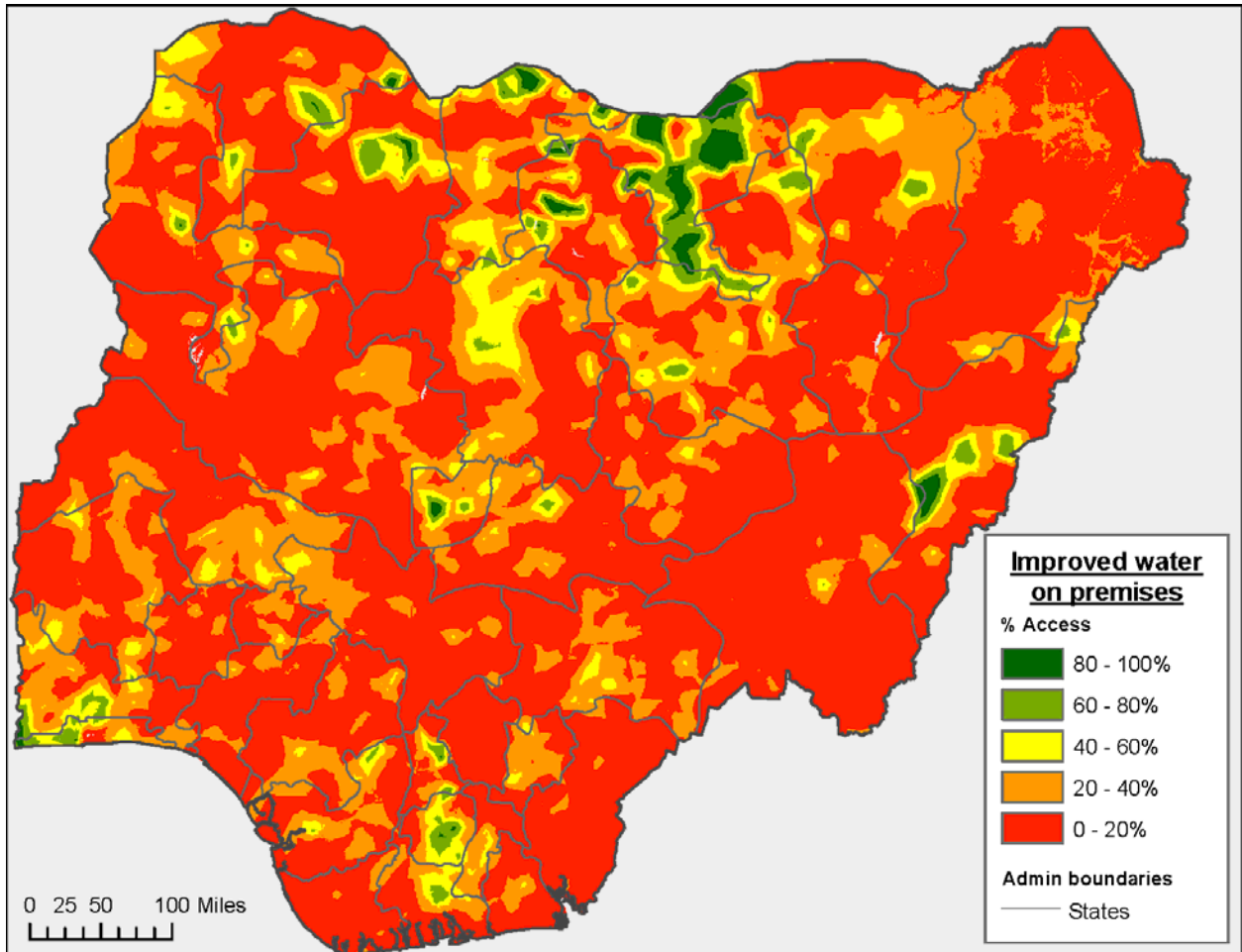
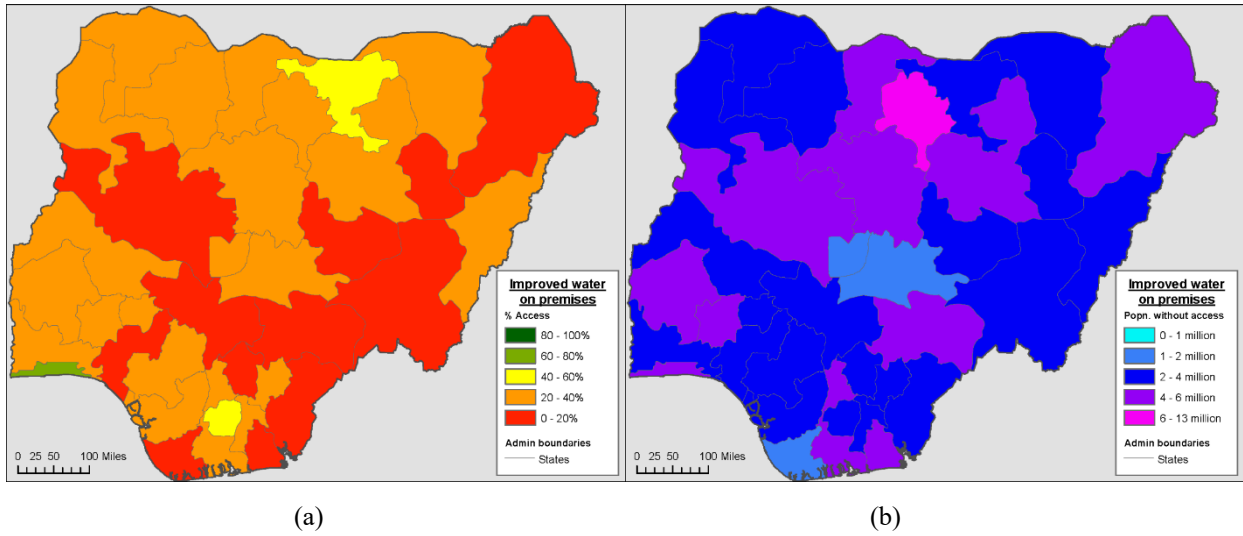


Figure 7 (main) Map showing modeled 1×1 km pixel level predictions of the percentage of population with improved water on the premises. Also shown are state-level estimates of (a) the percentage of people with improved water on premises and (b) the number of people without improved water on premises.

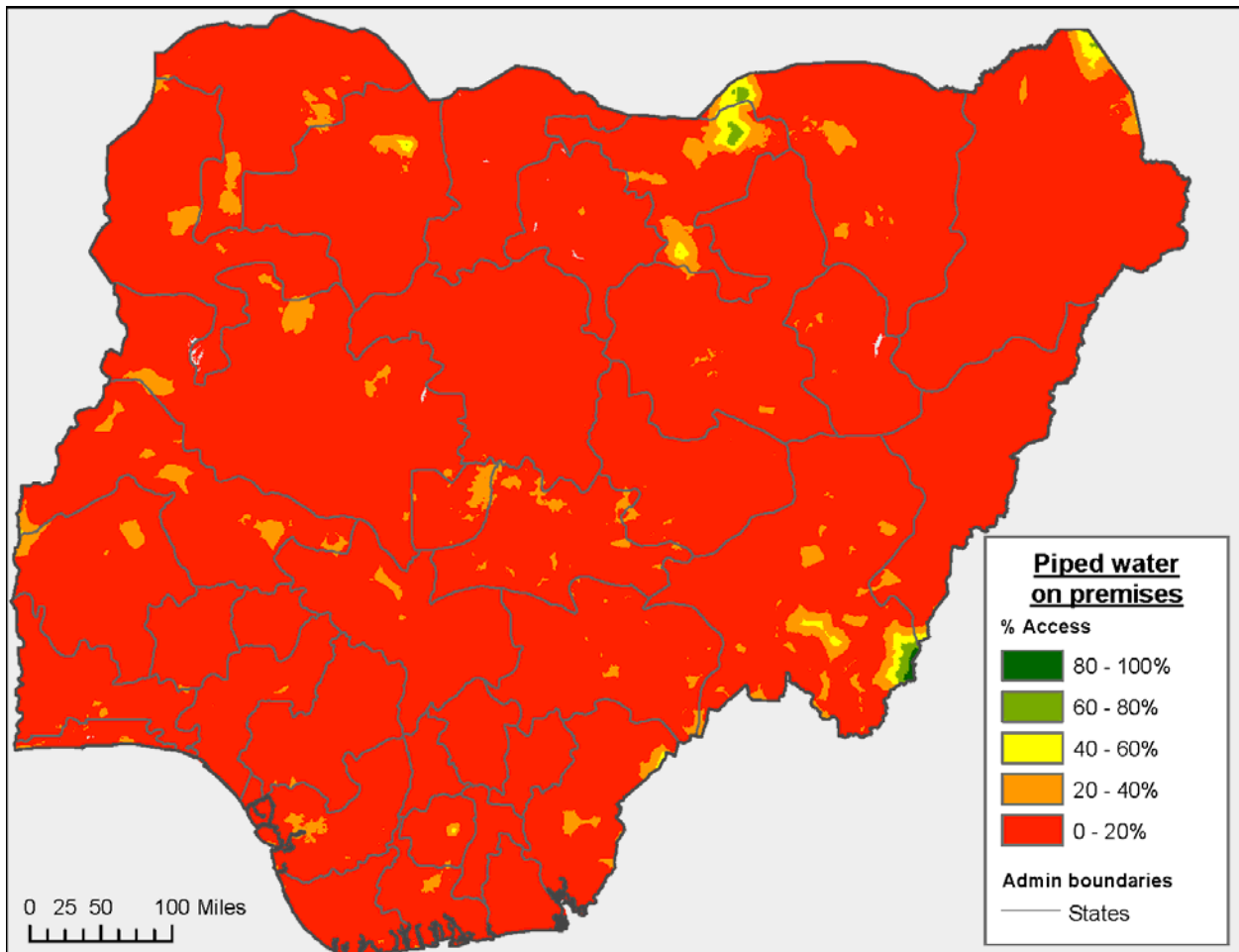
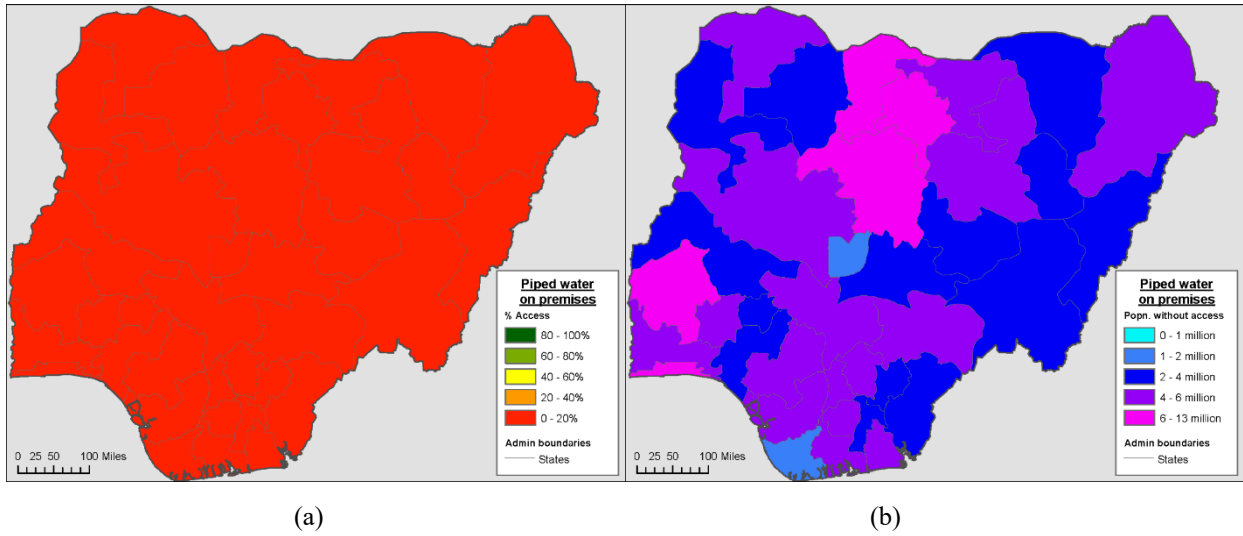


Figure 8 (main) Map showing modeled 1×1 km pixel level predictions of the percentage of the population with piped water on the premises. Also shown are state-level estimates of (a) the percentage of people with piped water on premises and (b) the number of people without piped water on premises.

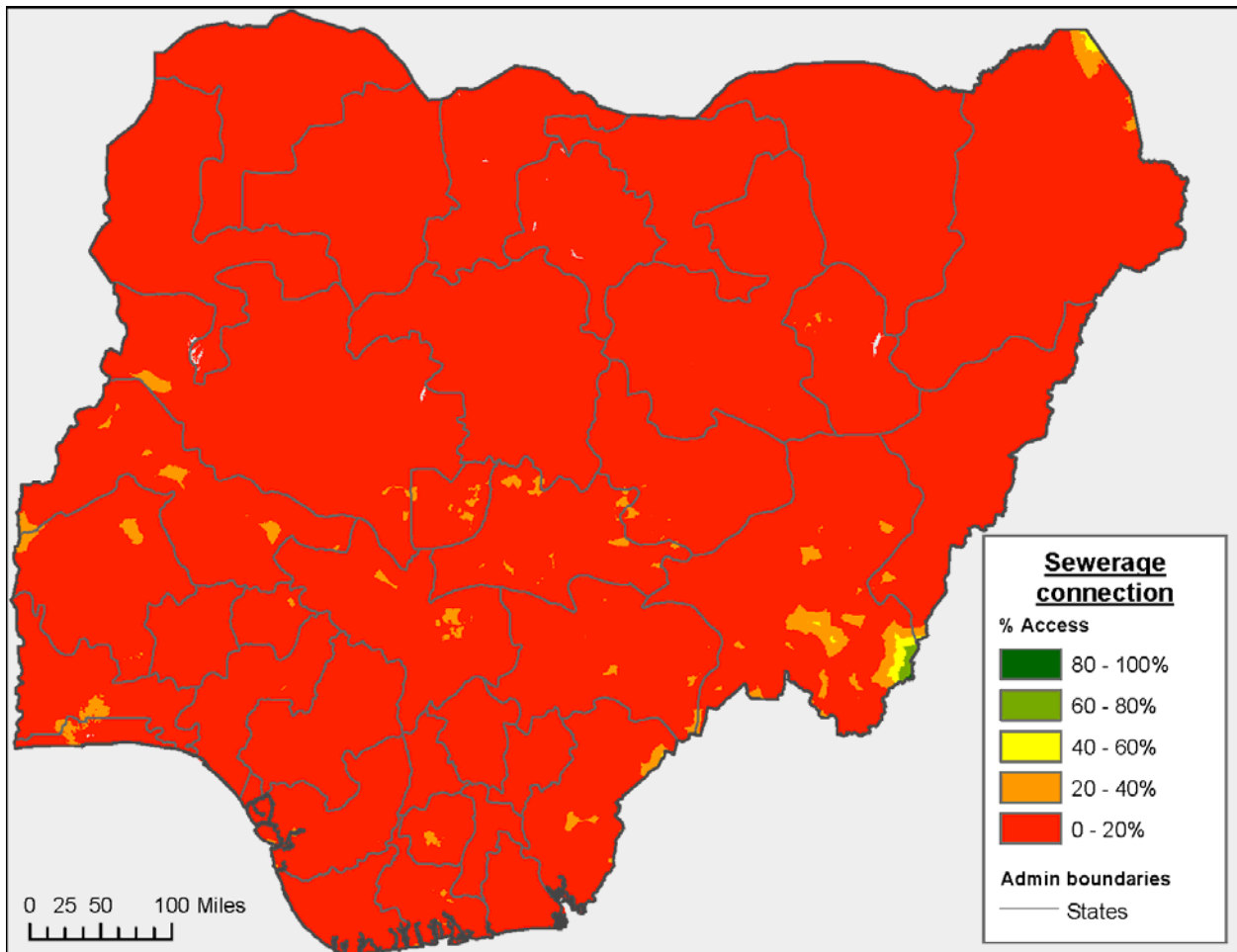
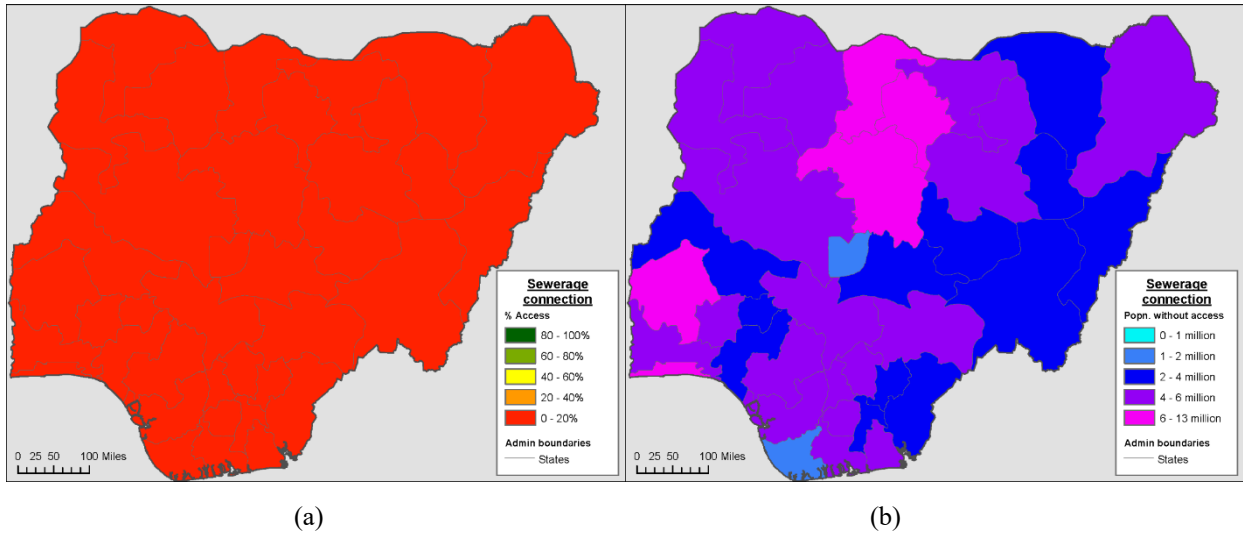


Figure 9 (main) Map showing modeled 1×1 km pixel level predictions of the percentage of population with a sewerage connection. Also shown are state-level estimates of (a) the percentage of people with a sewerage connection and (b) the number of people without a sewerage connection.

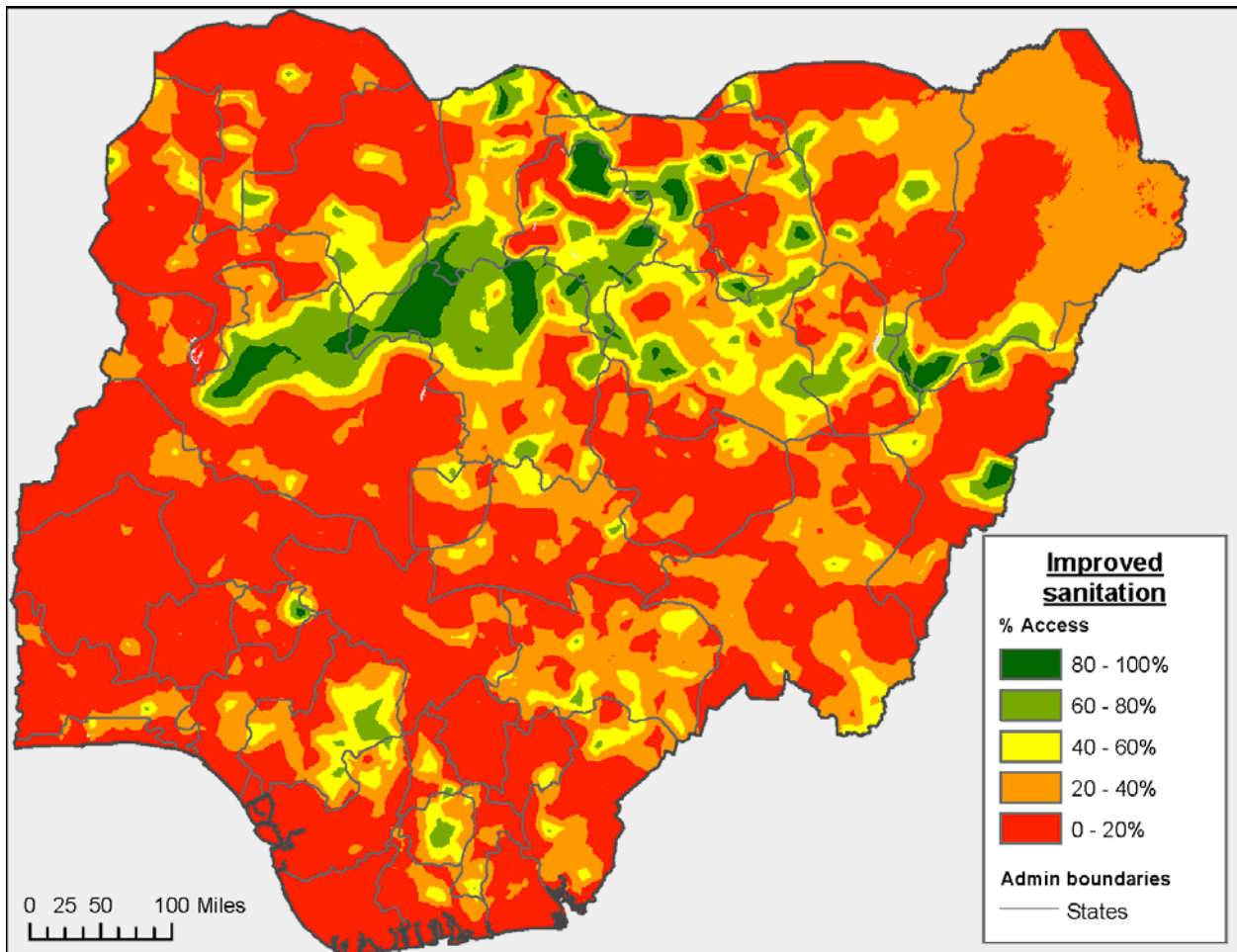
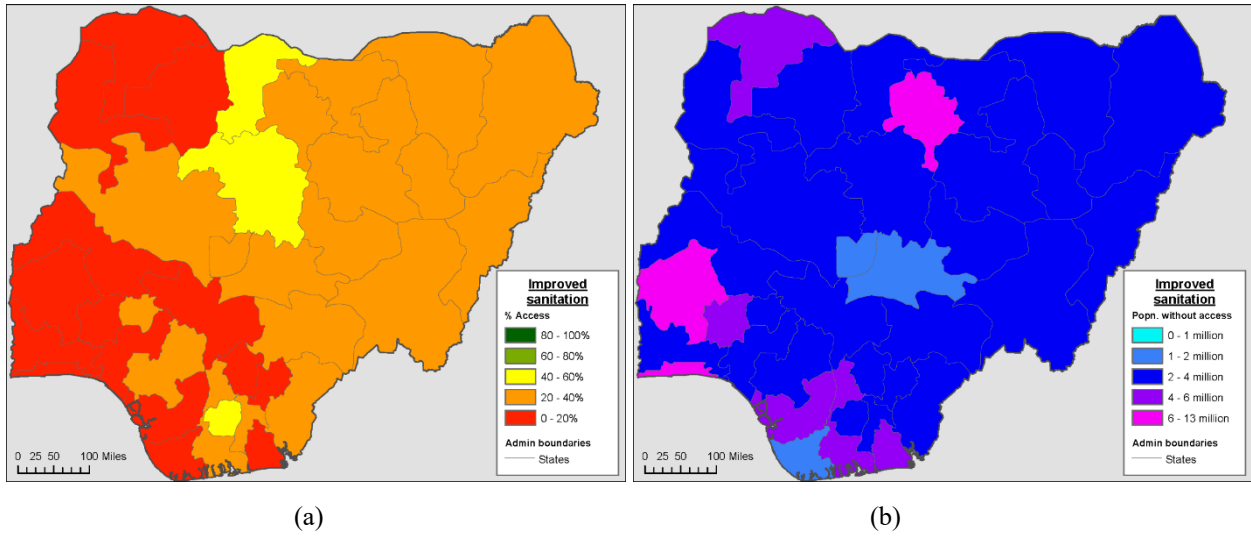


Figure 10 (main) Map showing modeled 1×1 km pixel level predictions of the percentage of the population with improved sanitation. Also shown are state-level estimates of (a) the percentage of people with improved sanitation and (b) the number of people without improved sanitation.

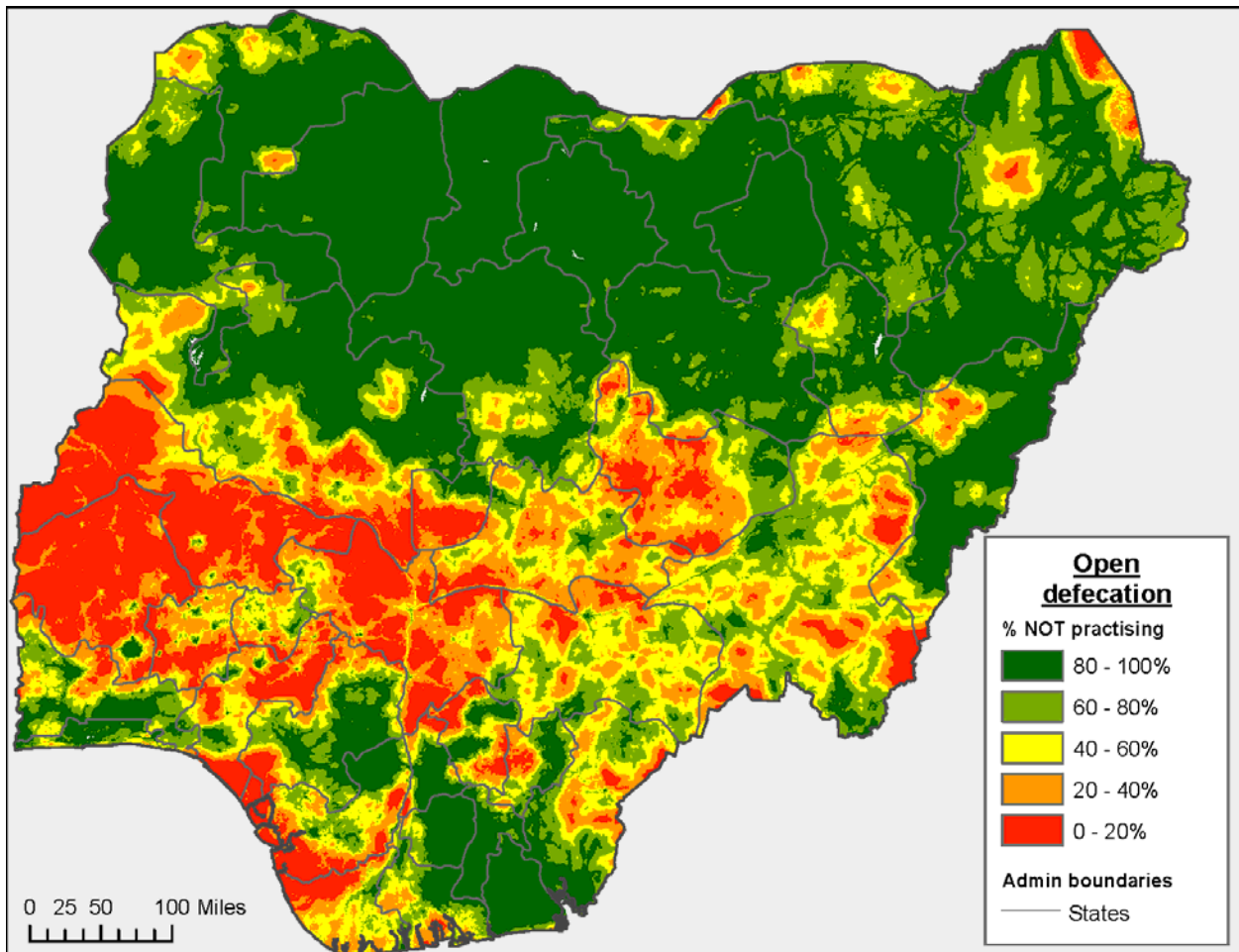
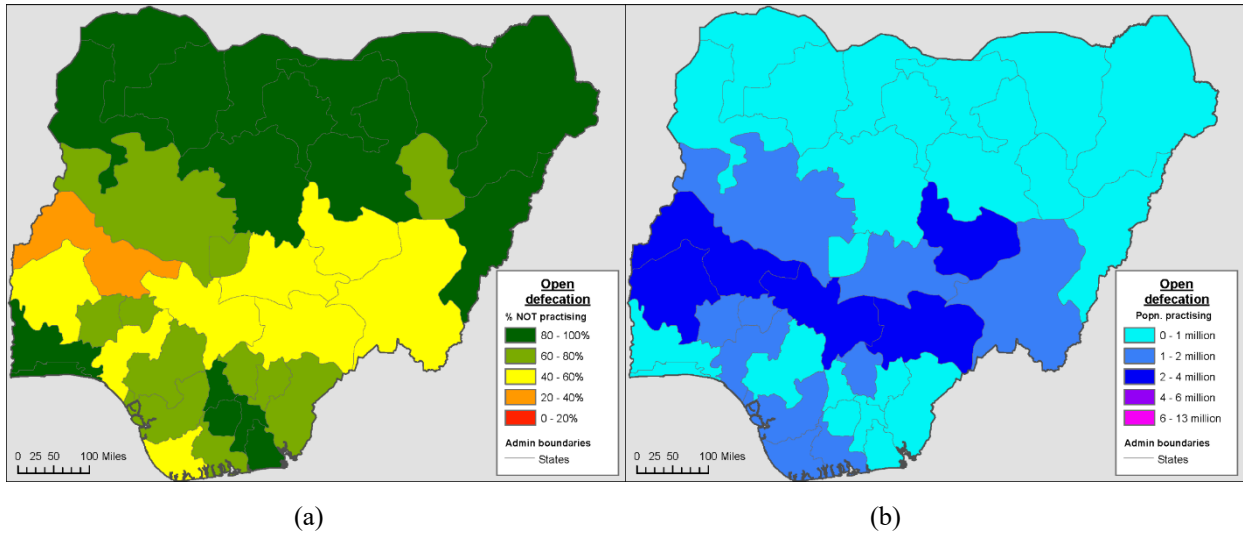


Figure 11 (main) Map showing modeled 1×1 km pixel level predictions of the percentage of population not practicing open defecation. Also shown are state-level estimates of (a) the percentage of people not practicing open defecation and (b) the number of people practicing open defecation.