Accepted Manuscript

A probabilistic predictive Bayesian approach for determining the representativeness of health and demographic surveillance networks

C. Edson Utazi, Sujit K. Sahu, Peter M. Atkinson, Natalia Tejedor, Andrew J. Tatem

PII: S2211-6753(16)30024-0
DOI: http://dx.doi.org/10.1016/j.spasta.2016.05.006
Reference: SPASTA 170

To appear in: Spatial Statistics

Received date: 29 January 2016
Accepted date: 13 May 2016

Please cite this article as: Utazi, C.E., Sahu, S.K., Atkinson, P.M., Tejedor, N., Tatem, A.J., A probabilistic predictive Bayesian approach for determining the representativeness of health and demographic surveillance networks. Spatial Statistics (2016), http://dx.doi.org/10.1016/j.spasta.2016.05.006

This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.
A probabilistic predictive Bayesian approach for determining the representativeness of health and demographic surveillance networks

C. Edson Utazi, Sujit K. Sahu, Peter M. Atkinson, Natalia Tejedor and Andrew J. Tatem

Abstract

Health and demographic surveillance systems, formed into networks of sites, are increasingly being established to circumvent unreliable national civil registration systems for estimates of mortality and its determinants in low income countries. Health outcomes, as measured by morbidity and mortality, generally correlate strongly with socioeconomic and environmental characteristics. Therefore, to enable comparison between sites, understand which sites can be grouped and where additional sites would aid understanding of rates and determinants, determining the environmental and socioeconomic representativeness of networks becomes important. This paper proposes a full Bayesian methodology for assessing current representativeness and consequently, identification of future sites, focusing on the INDEPTH network in sub-Saharan Africa as an example. Using socioeconomic and environmental data from the current network of 39 sites, we develop a multi-dimensional finite Gaussian mixture model for clustering the existing sites. Using the fitted model we obtain the posterior predictive probability distribution for cluster membership of each $1 \times 1$ km grid cell in Africa. The maximum of the posterior predictive probability distribution for each grid cell is proposed as the criterion for representativeness of the network for that particular grid cell. We demonstrate the conceptual superiority and practical appeal of the proposed Bayesian probabilistic method over previously applied deterministic clustering methods. As an example of the potential utility and application of the method, we also suggest optimal site selection methods for possible additions to the network.

Keywords: Bayesian Inference, BIC, Central clustering, Finite Gaussian mixture model, Gibbs sampling, Predictive clustering
1 Introduction

Health and demographic indices related to births, deaths, migration, economic activity, morbidity and infant mortality are used by governments and many other organizations for effective planning, health policy formulation and other decision-making processes (Ngom et al., 2001; WHO, 2013). The data used for estimating these indices are usually obtained from censuses and national civil registration systems. However, data from these sources in low-income countries are often inconsistent and unreliable (Sankoh and Byass, 2012; Ye et al., 2012). To circumvent this problem, longitudinal data collection systems known as Health and Demographic Surveillance Systems (HDSS) were established to provide more informative and accurate long-term monitoring data. Such systems involve the description of a target population through an initial census, which is then succeeded by regular collection of vital statistics and other relevant data. To provide a comprehensive picture of health and population dynamics across much wider geographical areas, many HDSS sites have joined networks such as the International Network of field sites with continuous Demographic Evaluations of Populations and Their Health (INDEPTH) (Baiden et al., 2006; Sankoh and Byass, 2012). At the time of writing, the INDEPTH network comprises of 52 HDSS sites, of which 39 are located in sub-Saharan Africa. Moreover, new networks are being established, such as the Child Health and Mortality Prevention Surveillance Network (CHAMPS), a network of disease surveillance sites in developing countries.

With networks of surveillance sites attempting to provide standardized and representative data on a range of health and demographic indicators, and pooling data to provide information on wide-area demographic patterns and their determinants, understanding the coverage and representativeness of the network becomes important. This is particularly necessary when it is of interest to determine where additions to the networks would help improve utility and coverage or to characterize the uncertainty associated with extrapolations using information from the network. Also, when performing site selection for establishing new networks of sites, understanding how the network can be effectively configured to capture the range of variabilities that exist in the regions of interest can be valuable.

Morbidity, mortality and health equity measures generally correlate and are influ-

\footnote{See \url{http://www.indepth-network.org/index.php?option=com_content&task=view&id=1306&Itemid=1070} for details.}

\footnote{See \url{http://www.gatesfoundation.org/Media-Center/Press-Releases/2015/05/Child-Health-and-Mortality-Prevention-Surveillance-Network}}
enced strongly by socioeconomic and environmental conditions. These have therefore previously been used as surrogate measures for assessing the coverage and representativeness of the INDEPTH network (Tatem et al., 2006; Jia et al., 2015). These studies used deterministic approaches such as hierarchical clustering (Ward, 1963) for grouping the sites based on gridded datasets depicting factors such as temperature, rainfall and population density, and the Euclidean metric for mapping the socioeconomic and environmental coverage of the sites. These studies provided a basic assessment of the similarities between existing sites, representativeness of the network and grouping of sites in terms of available gridded covariate layers, but did not account for or quantify the uncertainties inherent in undertaking this.

Model-based clustering methods offer a more statistically principled approach to clustering than heuristic methods such as hierarchical clustering. Model-based methods quantify the uncertainty in assigning observations to the clusters through a probability distribution. Finite mixture models (McLachlan and Peel, 2000; Fraley and Raftery, 2002) in which each mixing component corresponds to a cluster are well-known in this category. Inference in such models is concerned with estimating the model parameters and choosing the number of components to best describe the data. The models are estimated using maximum likelihood methods implemented using the expectation-maximization (EM) algorithm (Dempster et al., 1977) or Bayesian methods (Diebolt and Robert, 1994; Dey, Kuo, and Sahu, 1995). Bayesian approaches are more flexible and they allow for greater coherency in model estimation and, in particular, the predictive clustering proposed in this work. Also, during model estimation, uncertainty about the number of components can be incorporated in a Bayesian setting through a Dirichlet process prior (Escobar and West, 1995) or the use of a reversible jump MCMC sampler (Richardson and Green, 1997).

In this paper, we propose a probabilistic Bayesian approach to assess the representativeness of the INDEPTH HDSS network in sub-Saharan Africa as an alternative to previous approaches (Tatem et al., 2006; Jia et al., 2015). Our analyses are based on 1 × 1 km spatial resolution data for Africa using covariate layers representing eight variables described in Section 2. The probabilistic predictive approach is presented in two stages. First, we propose a multi-dimensional finite Gaussian mixture model fully implemented in a Bayesian framework for clustering the HDSS sites as discussed in Section 3. The Bayesian framework incorporates a recently developed central clustering procedure (see Mukhopadhyay et al., 2011) for summarizing the posterior distribution of the clustering configurations. This enables us to find the most representative
clustering configuration along with its uncertainty as provided by a 95% credible region. In the second stage, we perform predictive clustering of all the 1 × 1 km grid cells to determine which HDSS sites can realistically be clustered together in terms of the socioeconomic and environmental characteristics of the regions in which they are located. In the predictive clustering method, the grid cells are assigned to the clusters of the HDSS sites using their maximum predictive probabilities; see Section 4. Using the resulting probability map which clearly reveals inadequately covered areas, we employ a space-filling design to demonstrate how the location of new sites could be determined to help improve the coverage of the network, in terms of the covariate layers considered. Our analysis and the results are discussed in Section 5. A few summary remarks are placed in Section 6. An appendix contains details of the adopted prior distributions and also the details for the MCMC implementation.

2 Data description and exploratory analysis

Data for our analysis come from the 39 member sites of the African INDEPTH HDSS network at the time of writing; see Figure 3 for their locations. Table 1 provides the name, along with the country, population density and the size of the catchment area for each of these 39 sites. The table shows considerable variation in both the population density and the size and nature of the catchment area. The catchment areas of the sites were delineated through a combination of approaches. In some cases, the catchments were either digitized within ArcGIS from maps of the sites available on the INDEPTH website or obtained from the organizers of the network. In other cases where the catchments were administrative boundaries within countries, these were extracted from shapefiles obtained from national statistical offices or www.gadm.org.

Gridded socioeconomic and environmental data, for eight variables described in Table 2, are available for all 37 million 1 × 1 km grid cells covering all of the African continent. The first two socioeconomic variables, denoted by acc50k and gecon, measure respectively the travel time (in minutes) to nearest settlement of greater than 50k people and grid cell economic output in purchasing power parity in 2000 (Nordhaus, 2006). The third variable, popdens, is the population density per square kilometre measured in 2010 (Linard et al., 2012). The remaining five variables measure environmental characteristics and are given by: (i) alt, altitude above sea level (in metres), (ii) tmmpmean, annual mean temperature, (°C), (iii) pretot, annual total precipitation, (iv) t...
tmpseas, temperature seasonality as expressed by the coefficient of variation (100 \times standard deviation divided by the annual mean), (v) preseas, precipitation seasonality (coefficient of variation). For temperature and precipitation, the measurements are based on annual averages between 1950 and 2000. (More information on the methods used to obtain these environmental variables can be found in Hijmans et al. (2005).)

Each of these eight variables is called a layer, as in GIS terminology. The variables are plotted in Figure 1. These maps show obvious variation in the eight layers.

Given the focus on population health and demographics, unpopulated or lowly populated areas were masked out (the areas covered by the grid cells where the population density is less than one person per square kilometre). Hence, we work with the remaining 18,496,629, denoted by \( N \), grid cells. The ranges of gecon and popdens are very large (see Figure 1) and also they are positively skewed as is usually the case for population density and the socioeconomic indicator, gecon. Henceforth, we work with the logarithm of these variables which reduces skewness and encourages normality for the modelling proposed in Section 3. To avoid having to take logarithm of zero, we added 0.5 to gecon before applying the transformation.

There is still considerable variability between the values of the eight variables (after masking) even after taking the logarithm transformation. To eliminate the large layer specific variability we standardize all the variables to have unit variance simply by dividing each variable by the sample standard deviation of the \( N \) grid cell values. This allows us to conduct the representativeness study without having to deal with large numbers due to scale differences. Our mixture modelling effort and the Bayesian predictive probabilities are not affected by this change of scale because the scales of the layers are automatically modelled and taken care of in the probability calculation in Sections 3 and 4.

These pre-processed data for each of the eight layers for each of the 39 HDSS sites were extracted for all the 1 \times 1 km grid cells falling within the catchment area of the particular HDSS site. The summary statistics of the extracted data for the 39 sites are reported in Table 3. These summaries are unit-free due to the standardization performed in the previous paragraph.

The catchment area of each of the 39 sites contains multiple 1 \times 1 km grid cells (see the last column of Table 1) which poses a potential problem of spatial misalignment while performing the comparison between each site and an arbitrary 1 \times 1 km grid cell from the total \( N \). To alleviate this problem, we simply average the values of each of the 8 layers for all the 1 \times 1 km grid cells falling within the catchment so that a unique
value for each layer was obtained for each of the 39 sites. We studied the sensitivity of this choice by taking the median and the mode of the catchment area grid cells both of which yielded the same clustering as the averages. This analysis assures us that the results are not affected unduly by our choice of the average as representative of the catchment area grid cells. Moreover, the misalignment problem disappears since the median corresponds to a particular 1 × 1 km grid cell within the site catchment area and the clustering predictions are performed for 1 × 1 km grid cells.

3 Bayesian model-based clustering methods

3.1 The Bayesian finite Gaussian mixture model

Let \( y_i \in \mathbb{R}^d \) denote the \( d \) average measurements from the \( i \)th HDSS site, \( i = 1, \ldots, n = 39 \). To group these \( n \ d \)-dimensional observations into an unknown number of clusters, we seek an appropriate model. Here none of the components of \( y_i \) stands out as a natural dependent variable, hence a typical regression modelling setup is not applicable. This is why, we treat each \( y_i \) as a multi-dimensional random variable assumed to follow the Gaussian mixture model with \( K \) components:

\[
y_i | \Theta_K \sim \sum_{j=1}^{K} \pi_j N_d(y_i; \theta_j), \quad i = 1, \ldots, n;
\]

where \( \Theta_K = (\pi_1, \ldots, \pi_K, \theta_1, \ldots, \theta_K) \), \( \theta_j = (\mu_j, \Sigma_j) \), and the parameters \( \mu_j, \Sigma_j \) and \( \pi_j (\pi_j \in (0, 1) \) and \( \sum_{j=1}^{K} \pi_j = 1) \) denote the mean vector, the covariance matrix and the mixing proportion of the \( j \)th component respectively.

Inference using (1) is often facilitated by its missing data representation. Let the categorical random variables \( z_i, i = 1, \ldots, n \), denote the cluster index of the \( i \)th site. That is, \( z_i = j \) if the \( i \)th site belongs to the \( j \)th cluster \( (j = 1, \ldots, K) \). For future reference, let \( z = (z_1, \ldots, z_n) \) denote the unknown cluster indicators and \( y = (y_1, \ldots, y_n) \) denote the full set of data from the 39 HDSS sites. The model can now be expressed in a hierarchical form as

\[
y_i | z_i = j, \Theta_K \sim N_d(y_i; \theta_j);
\]

\[
p(z_i = j | \Theta_K) = \pi_j.
\]

In the model, each component is characterized by its mean vector \( \mu_j \) and covariance
matrix $\mathbf{\Sigma}_j$. The $\mu_j$'s form the centers of the clusters whereas the $\mathbf{\Sigma}_j$'s determine the geometric features of the clusters; see Fraley and Raftery (2007). Each $\mathbf{\Sigma}_j$ in the model contains $d(d+1)/2$ parameters. With large and even moderate $d$ as seen in Section 2, it is often the case that we do not have enough data to obtain reliable estimates of these parameters for each component. Small $n$ and moderate or large $d$ would usually lead to a distortion of the geometric attributes of the clusters, culminating in wrong estimation of the number of components and misclassification of the data. Hence, to preserve parsimony in the model, different parameterizations of the covariance matrix describing varying geometric attributes - shape, size and orientation - of the clusters in the Euclidean space could be considered (Fraley and Raftery, 2007; Bouveyron and Brunet-Saumard, 2014). These parameterizations can also be viewed as constrained versions of the full model in (2), with the constraints designed to reduce the number of parameters to be estimated. In this work, we consider the following parameterizations:

$$
\mathcal{M}_1: \quad \mathbf{\Sigma}_j = \sigma^2 \mathbf{I} \quad \forall \ j; \quad (3)
$$
$$
\mathcal{M}_2: \quad \mathbf{\Sigma}_j = \mathbf{\Sigma} \quad \forall \ j; \quad (4)
$$
$$
\mathcal{M}_3: \quad \mathbf{\Sigma}_j = \text{diag}(\sigma_{1j}^2, \ldots, \sigma_{dj}^2) \quad \forall \ j; \quad (5)
$$
$$
\mathcal{M}_4: \quad \mathbf{\Sigma}_j = \sigma_j^2 \mathbf{I}; \quad (6)
$$
$$
\mathcal{M}_5: \quad \mathbf{\Sigma}_j = \text{diag}(\sigma_{1j}^2, \ldots, \sigma_{dj}^2); \quad (7)
$$
$$
\mathcal{M}_6: \quad \mathbf{\Sigma}_j = \mathbf{\Sigma}_j; \quad (8)
$$

with model $\mathcal{M}_6$ in (8) being the full unconstrained model specified in (2). Model $\mathcal{M}_1$ as well as models $\mathcal{M}_3$ to $\mathcal{M}_5$ constrain the covariance matrix to be diagonal and thus assume that the variables are conditionally independent. In model (4), the mixture components are assumed to have the same covariance matrix. Clearly, these constrained models reduce the number of covariance parameters to be estimated significantly. Several other model parameterizations based on the eigenvalue decomposition of $\mathbf{\Sigma}_j$ as given in Banfield and Raftery (1993) and Celeux and Govaert (1995) are also possible.

The Bayesian model specification in each case of the full and constrained models must be completed by assuming prior distributions for all the unknown parameters. Details of these prior distributions and the implementation of the models including the MCMC algorithms are provided in the Appendix.
3.2 Choosing the best model

To fit the finite mixture model in (2), it is necessary to specify a suitable method for choosing the number of clusters as well as determining the best model parameterization. In the statistical literature, many approaches have been proposed for dealing with model choice in a Bayesian framework. These include: Bayesian information criterion (BIC) (Schwarz, 1978), Bayes factors (Kass and Raftery, 1995), deviance information criterion (DIC) (Spiegelhalter et al., 2002) and the distance-based method of Sahu and Cheng (2003). In the analysis of finite Gaussian mixture models, Steele and Raftery (2009) compared the performance of full and empirical Bayesian methods for selecting the number of components in mixture models. They found that BIC outperformed all the other methods they considered. Based on their findings, we use the BIC - an approximation to the Bayes factor - to determine the best model as well as the best $K$.

The modified version of the BIC used is defined as

$$BIC = -2 \log p(\Theta_{M,K}; y) + \gamma_K \log n,$$

(9)

where $p(.)$ is the ‘complete-data’ likelihood of model $M$ with $K$ components evaluated using the maximum a posteriori (MAP) estimate, $n$ is the sample size and $\gamma_K$ is the number of free parameters in the model, which for model (8) is $(K - 1) + Kd + Kd(d + 1)/2$. In our MCMC sampling scheme with $T$ samples from the posterior distribution, we choose $\hat{\Theta}_{M,K}$ to be the MAP estimate given by:

$$\hat{\Theta}_{M,K} \approx \arg \max_{\Theta_{M,K}; 1 \leq t \leq T} \left\{ p(\Theta_{M,K}^{(t)}; y) \times p(\Theta_{M,K}^{(t)}) \right\},$$

(10)

where $p(\Theta_{M,K})$ denotes the prior distribution for $\Theta_{M,K}$. The best fitting model and number of clusters both have the minimum BIC value, which we denote by $k^*$. We found that this BIC specification performed well in simulation studies for selecting the best model and $K$, outperforming the DIC (and even its modifications - see Celeux et al. (2006)) which under-penalizes complex models.

3.3 Central clustering

Once the number of clusters, $k^*$, corresponding to the best model, is chosen by the BIC as detailed above, we aim to identify the most representative clustering config-
uration based on the recently developed idea of central clustering (see, for example, Mukhopadhyay et al., 2011) as follows.

Each iteration of the MCMC algorithm yields a particular clustering of the HDSS sites, even when the number of clusters remain unchanged between iterations. When \( T \) draws are obtained, a suitable method for summarizing the posterior distribution of the clustering configurations, \( z_1, \ldots, z_T \), is required. In so doing, we aim to obtain the clustering that is “central” and most representative of all the clusterings obtained alongside its corresponding credible region. Empirically, \( z^* \) is a central clustering if for a given small \( \epsilon > 0 \) (typically, \( 0 < \epsilon < 1 \)),

\[
    z^* = \arg \max_{1 \leq l \leq T} \frac{1}{T} \# \{ z_l; 1 \leq l \leq T : d(z_t, z_l) < \epsilon \},
\]  

(11)

where \( d(\cdot, \cdot) \) is a suitably chosen metric that measures the dissimilarity or otherwise of a pair of clusterings and \( \#\{A\} \) denotes the cardinality of the set \( A \). Note that for the multimodal posterior distribution of clusterings, varying \( \epsilon \) over \((0, 1)\) and letting \( \epsilon \to 0 \) will lead to the detection of all the modes and the global mode.

Let \( n_{ij} \) denote the number of observations in the \( i \)th cluster of \( z_l \) and \( j \)th cluster of \( z_t \) for any \( l \) and \( t \), \( 1 \leq l, t \leq T \). The \( n_{ij} \)'s are easily determined from a \( k^* \times k^* \) cross-tabulation of the clusterings. An approximate metric for determining the distance between clusterings \( z_l \) and \( z_t \) is given by

\[
    d(z_l, z_t) = \max \left\{ \tilde{d}(z_l, z_t), \tilde{d}(z_t, z_l) \right\},
\]

(12)

where \( \tilde{d}(z_l, z_t) = 1 - \frac{\max_{1 \leq i \leq k^*, 1 \leq j \leq k^*} n_{ij}}{n_{00}} \) with \( n_{00} = \sum_{i=1}^{k^*} \sum_{j=1}^{k^*} n_{ij} \). An approximate 95% credible region for the central clustering \( z^* \) is defined as the set \( \{ z_t; 1 \leq t \leq T : d(z_t, z^*) < \epsilon^* \} \), where the quantity \( \epsilon^* \) is such that

\[
    \frac{1}{T} \# \{ z_t; 1 \leq t \leq T : d(z_t, z^*) < \epsilon^* \} \approx 0.95.
\]

(13)

This credible region is constructed by initially setting \( \epsilon^* = 0 \) and then successively adding a very small quantity (e.g. \( 10^{-10} \)) until (13) is satisfied. When there are multiple modes (i.e. more than one central clustering), a highest posterior density region can also be constructed adaptively. Further details about this concept including theoretical properties are given in Mukhopadhyay et al. (2011).
4 Probabilistic predictive clustering and site selection

4.1 Cluster representativeness via predictive clustering

To determine the spatial coverage or representativeness of the HDSS sites, we now detail the method for performing predictive clustering of the $1 \times 1$ km grid cells. Let $y_m$ denote the 8-dimensional socioeconomic and environmental measurements for the $m$th $1 \times 1$ km grid cell, $m = 1, \ldots, N (= 18,496,629)$. Let $z_m$ denote the unknown cluster index for $y_m$. Our objective here is to find, $p(z_m = j | y, y_m, k^*)$ for $j = 1, \ldots, k^*$ where $k^*$ is the chosen value of the optimum number of clusters. This probability can be expressed as

$$ p(z_m = j | y, y_m, k^*) = \int p(z_m = j | y_m, \Theta_{k^*}, z) p(\Theta_{k^*}, z | y, y_m) d\Theta_{k^*} dz, $$

where $\Theta_{k^*}$ contains the parameters of the chosen, $k^*$-component model. From (14), it can be seen that the inclusion of an additional data (grid cell) will change the posterior distribution of the parameters. This requires that the MCMC sampler (or, in our case, the Gibbs sampler) is rerun with all the $y_m$’s to be classified. This is infeasible and computationally prohibitive given the massive number of predictions to be made. As a solution to this problem, we propose a retrospective prediction strategy which involves approximating the integral in (14) as follows:

$$ p(z_m = j | y, y_m, k^*) \approx \int p(z_m = j | y_m, \Theta_{k^*}, z) p(\Theta_{k^*}, z | y, y_m) d\Theta_{k^*} dz, $$

$$ = E_{\Theta_{k^*}, z} (p(z_m = j | y_m, \Theta_{k^*}, z)). $$

Equation (15), also known as the posterior predictive distribution of $z_m$, can then be computed using the MCMC estimates of $(\Theta_{k^*}, z)$ by taking the empirical average of

$$ p(z_m = j | y_m, \Theta_{k^*}, z) \approx \frac{\pi_j \mathcal{N}_d(y_m; \theta_j)}{\sum_{l=1}^{k^*} \pi_l \mathcal{N}_d(y_m; \theta_l)}, \quad j = 1, \ldots, k^*, $$

over all the MCMC samples (see Richardson and Green (1997) for a related discussion). Given the probabilities from (16), the cluster membership of the $m$th grid cell is given by

$$ \hat{z}_m = \arg\max_j \{ p(z_m = j | y, y_m, k^*) \}. $$
Thus, each grid cell is assigned to the cluster in which it has the maximum probability. The maximum probabilities used in clustering the grid cells provide insights as to the representativeness of the clusters of the HDSS sites. Grid cells in poorly represented areas will be equally likely to be classified into any of the $k^*$ clusters. To identify these areas, we simply search for grid cells whose maximum classification probabilities are $\approx 1/k^*$. In general, and depending on the application at hand, a threshold classification probability can be established for the identification of poorly covered areas. These areas will further constitute candidate locations for the establishment of new sites to help maximize the socioeconomic and environmental coverage of the network.

4.2 Sampling design for site selection

We now describe a spatial sampling design to demonstrate the utility of the classification probabilities for the establishment of new sites using poorly covered areas as candidate locations.

Let $N_0 = \#\{m : \max\{p(z_m = j|y, y_m, k^*)\} < p_0; 1 \leq m \leq N, 1 \leq j \leq k^*\}$ denote the number of grid cells whose classification probabilities were less than the threshold probability, $p_0$. Let $D = \{s_1, \ldots, s_{N_0}\}$ denote the spatial (longitude and latitude) coordinates of the $N_0$ grid cells. To select $n_0$ candidate sites whilst keeping the current $n$ sites fixed, we use the space-filling design algorithm of Royle and Nychka (1998) as implemented in the fields package in R (R Core Team, 2013). The algorithm seeks to minimize the coverage criterion

$$Q(\tilde{D}, D) = \left( \sum_{s_i \in D} \left( \sum_{\tilde{s}_i \in \tilde{D}} \omega(s_i, \tilde{s}_i)^{r_1} \right)^{r_2/r_1} \right)^{1/r_2}; \quad (18)$$

where $\omega(\cdot, \cdot)$ denotes a spatial distance measure (e.g. the geodetic distance used here) of its arguments, and the parameters $r_1 < 0$ and $r_2 > 0$ are suitably chosen to yield a reasonable design, the set of $n + n_0$ locations in $\tilde{D}$ represent the design set optimized by the algorithm. See Royle and Nychka (1998) for details of the algorithm.

In this sampling design, certain eligibility conditions can be straightforwardly imposed to further restrain the set of candidate sites. For example, if grid cells in poorly covered areas whose population densities are below a pre-specified value are not to be considered, these can be masked out before the site selection is performed.
5 Data analysis

In this section, we discuss the application of the methodology to determine the socio-economic and environmental representativeness of the 39 INDEPTH HDSS network in sub-Saharan Africa. Implementation details are as provided in the Appendix. Table 4 reports the BIC values of models $M_1$ to $M_6$ for the given values of $K$. The BIC values are also plotted in Figure 2. It can be seen that the constrained models, $M_1$ to $M_5$, are clearly better than the full model, $M_6$ according to the BIC. Further, models $M_2$ and $M_3$ with the same covariance matrix for all the components fit the data better than other constrained models. The best fitting model is $M_3$ (BIC = 643.51) with $k^* = 5$ components.

From the application of the central clustering methodology of Section 3.3 to the posterior distribution of the clustering configurations, we obtained one central clustering ($z^*$) provided in Table 5. The value of $\epsilon$ corresponding to $z^*$ is 0.05. (This $\epsilon$ value was determined as the center of the distances between the clusterings. We note that for other values within $\pm 0.03$ of this $\epsilon$ value and for $\epsilon \to 0$, no other central clustering was detected.) Figure 3 plots the central clustering reported in Table 5. It can be seen that Cluster 2 is the largest cluster with 13 sites, Clusters 1 and 3 each have 7 sites while each of Clusters 4 and 5 contain 6 sites. The figure shows that two clusters - Clusters 2 and 3 - were identified in Western Africa. Interestingly, Cluster 3 is also made up of some sites in East Africa. This shows that these groups of sites are similar in characteristics although they are located in different regions of the continent. The other two clusters identified in East Africa are Clusters 1 and 4. All the sites in Southern Africa are shown to be similar. The 95% credible region for the central clustering, which also happens to the 95% highest posterior density region due to unimodality, was obtained as $\{z : d(z^*, z) < 0.1282\}$. The empirical probability of the credible region is 0.9608, which indicates a close approximation.

Comparison with $K$-means and hierarchical clustering: To demonstrate the superiority of the Bayesian clustering approach (using a finite Gaussian mixture model) with regards to clustering of the HDSS sites and accounting for the associated uncertainty, the $K$-means (MacQueen, 1967) and hierarchical clustering (with complete linkage method) algorithms were used to cluster the sites. In both analyses, we set the number of clusters equal to that of the central clustering as these methods offer no principled approach for determining the number of clusters. This meant that for hierarchical clustering, the dendrogram was cut at a height that yielded five clusters.
The plots of the clusterings in the space of the first two principal components of the data are shown in Figure 4. (Note that the clusterings do not have the same numbering.) These plots show clearly the variation among these clustering methods. Using the dissimilarity measure in (12), the distance between the \( K \)-means clustering and the central clustering was found to be 0.1026. Coincidentally, the distance between hierarchical clustering and the central clustering was also 0.1026. (The distance between \( K \)-means and hierarchical clustering was 0.0513.) This shows that the central clustering is close to the clusterings obtained using the \( K \)-means and hierarchical methods with both clusterings falling within its 95% credible region. We note that the central clustering was obtained taking into account the uncertainty regarding the clustering of the data and, hence, is more representative of other possible clusterings.

To assess the spatial coverage/representativeness of the clusters of the HDSS sites, we extrapolated the clusterings obtained to the \( 1 \times 1 \) km grid cells across Africa as discussed in Section 4.1. The results of this analysis are displayed in Figures 5 and 6. From the cluster map in Figure 5, it can be seen that the sites in Cluster 2 cover mainly the Sahel and other arid regions of the continent, where climatic and environmental conditions often show significant seasonal fluctuations (see Tatem \textit{et al.}, 2006). Cluster 3 generally aligns with the highly vegetated, tropical regions of the continent whereas Clusters 1 and 4 correspond more to the highland areas - the demographic and environmental characteristics of which are visible from examining Figures 1c - 1f. Lastly, the sites in Cluster 5 typically characterize the steppes and semi-desert regions of the continent.

The probability map in Figure 6 displays the maximum probabilities (see equation (17)) used in clustering the grid cells. The map depicts the uncertainty in the predictive clustering of the grid cells based on the clustering of the HDSS sites. A high probability is indicative of lower uncertainty while a low probability suggests greater uncertainty. The light coloured areas are well represented by the network and are classified with high probabilities. Poorly represented areas are dark coloured and these are classified with low probabilities. For example, the region bordering Central and Southern Africa is poorly represented by the current network. The lowest probability of classification was obtained as 0.23. It should be noted that grid cells with classification probabilities of \( \approx 0.2 \) are equally likely to belong to any of the five clusters. Hence, these areas are the most poorly represented by the network.

Finally, with a threshold probability value of \( p_0 = 0.4 \), the spatial sampling design described in Section 4.2 was applied to choose the location of \( n_0 = 10 \) new sites. This
choice of $p_0$ implies that grid cells whose classification probabilities were less than 0.4 were not considered as candidate sites. Following test runs, we set $r_1 = -100$ and $r_2 = 1$. The implementation of the algorithm produced the coverage design shown in Figure 6. The green filled circles in the figure are locations where the establishment of new sites will help maximize the socioeconomic and environmental coverage of the network, given the socioeconomic and environmental layers considered here. Hence, the methodology proposed here offers an integrated approach that produces outputs both for determining the representativeness of the current network and a template for improving its coverage.

6 Summary and discussion

We have proposed and applied a Bayesian methodology for assessing the coverage of a network of health and demographic surveillance sites using gridded socioeconomic and environmental data. Our analysis of the coverage of the INDEPTH HDSS network in sub-Saharan Africa built upon the work of Tatem et al. (2006) and Jia et al. (2015) to map the sociodemographic and environmental coverage of the INDEPTH HDSS network in sub-Saharan Africa at $1 \times 1$ km spatial resolution. The central clustering of the HDSS sites, the cluster and probability maps resulting from predictive clustering of the $1 \times 1$ km grids and the illustrative location of new sites comprise the key outputs from the application of the methodology. These outputs have demonstrated clearly the representativeness of the network through the identification of poorly and well covered areas, among other important findings.

In Tatem et al. (2006) and Jia et al. (2015), separate analyses were carried out using deterministic approaches to cluster the sites, to perform environmental and socioeconomic classification of the grid cells and finally, to determine the representativeness of the sites. However, the Bayesian methodology developed in this paper offers a more coherent approach in that it allows us to cluster the sites, to characterise site representativeness and to undertake a socioeconomic and environmental classification of the entire continent of Africa all in a single model-based analysis. Furthermore, considering the large number of $1 \times 1$ km grid cells in Africa, it will be computationally prohibitive (if at all possible) to cluster these using deterministic approaches such as the $K$-means and hierarchical clustering when using statistical software packages such as R. However, with our methodology, the grid cells can be easily clustered by post-MCMC parallel computing on a high-performance computer.
Lack of ready to use packaged software routines is a potential weakness of the proposed method. The intricacy of our methodology meant that currently, the clustering of the sites using mixture modelling, model choice using BIC, determination of the central clustering, predictive clustering of the 1 x 1 km grid cells and site selection are each executed using separate computer code. Moreover, the choice of $\epsilon$ for determining the central clustering is iterative and is currently performed by trial and error. Thus, the methodology as a whole is intensive requiring decision making at several stages. Hence, to facilitate routine implementation of the methodology further investment is required in flexible software development to integrate the component sub-procedures, which we plan to take up elsewhere.

Further work on the methodological aspects will focus on exploring a variable dimensional MCMC method (Richardson and Green, 1997) for clustering the sites. Also, extensions to a nonparametric mixture model using a Dirichlet process prior may be worth investigating. Nevertheless, considering the nature of the data analysed, we note that such extensions should be handled with caution to avoid over-fitting problems. Over-parameterizing the model can result in noisy parameter estimates and cause problems during predictive clustering of the grid cells. Therefore, any extension incorporating the uncertainty about the number of components via the aforementioned approaches should also involve parsimonious parameterizations of the covariance matrices of the mixture components.

The distance-based sampling design discussed in Section 4.2 for the selection of new sites is an exploratory approach. Extensions, in line with the rationale behind the establishment of HDSS networks, will consider model-based approaches for sites selection (see, for example, Chipeta et al., 2016). These model-based approaches could be used to optimize the prediction of some health outcomes of interest such as under-five mortality and malaria prevalence over the poorly covered areas as obtained using the maximum classification probabilities.

As shown in Table 1, the catchment areas of the HDSS sites vary greatly in size. The implication of this is that within a given site, the values of a particular variable may exhibit a marked variability. For example, within Rakai, there exist clusters of grid cells with very different population densities. This information is lost when the sites are clustered based on the averages of the variables within their catchment areas. An extension of this work will therefore seek to incorporate within-site variability when clustering the sites. This could be in the form of an extension of model (1) to allow clustering of the sites using all the grid cells within their catchment areas.
Moreover, environmental conditions vary month by month, meaning that cluster structures will likely also vary monthly, and future work will explore the effects of this seasonality through monthly variables (see, for example, Dash et al., 2010). Further, additional variables will be explored, such as those relating to population nutrition and sanitation, which can have important effects on population health. Such variables may only be available as administrative unit-level aggregates, and therefore will require the development of methods for their integration into the grid-based approaches outlined in this paper.

References


Appendix: posterior details and the Gibbs sampler

Given the set of the measurements of the HDSS sites and their cluster indicators \( \{(y_i, z_i); i = 1, \ldots, n\} \), the ‘complete-data’ likelihood of the model is

\[
p(\Theta_K; y, z) = \prod_{j=1}^{K} \prod_{i: z_i = j} \pi_j \mathcal{N}_d(y_i; \theta_j),
\]

from which the observed-data likelihood can be easily obtained as \( p(\Theta_K; y) = \int p(\Theta_K; y, z)dz \) if desired. In a Bayesian setting, appropriate prior distributions are placed on the parameters. Assuming prior independence, we choose conjugate priors for the...
parameters as follows \((j = 1, \ldots, K; r = 1, \ldots, d)\):

\[
\pi_1, \ldots, \pi_K \sim \text{Dir}(\delta_1, \ldots, \delta_K); \\
\mu_j | \Sigma_j \sim N_d(\mu_a, 1/b \Sigma_j); \\
\Sigma_j, \Sigma \sim \text{Inverse Wishart}(s, S); \\
\sigma^2, \sigma^2_j, \sigma^2_r, \sigma^2_r \sim \text{Inverse Gamma}(s/2, S/2). \tag{a.2}
\]

That is, the mixing weights, the mean vectors, the covariance matrices and other covariance parameters are assumed to follow the Dirichlet distribution, the multivariate normal distribution, the inverse Wishart distribution and the inverse gamma distribution respectively.

For each model, the posterior distribution is given by \(p(\Theta_K; y, z) \times p(\Theta_K)\), which is simply the product of the likelihood in (a.1) and the joint prior distribution on the parameters, \(p(\Theta_K)\). It is straightforward to derive the conditional posterior distribution of the parameters for Gibbs sampling. These are provided as follows.

By Bayes’ theorem, the conditional posterior distribution of the indicator variables, \(z_i (i = 1, \ldots, n)\), in all the models can be obtained as

\[
p(z_i = j | y, \Theta_K, z_{-i}) = \frac{\pi_j N_d(y_i; \theta_j)}{\sum_{i=1}^K \pi_i N_d(y_i; \theta_i)}; \quad j = 1, \ldots, K. \tag{a.3}
\]

Let \(n_j = \# \{i : z_i = j\}\) and \(\tilde{y}_j = \sum_{i: z_i = j} y_i / n_j\). The conditional posterior distribution of the mixing weights \(\pi = (\pi_1, \ldots, \pi_K)\) is given by

\[
\pi | z \sim \text{Dir}(n_1 + \delta_1, \ldots, n_K + \delta_K). \tag{a.4}
\]

For the component means, \(\mu_j\), we have

\[
\mu_j | y, z, \Sigma_j \sim N_d(\tilde{\mu}_a, 1/b \tilde{\Sigma}_j); \tag{a.5}
\]

where \(\tilde{\mu}_a = (n_j \tilde{y}_j + b \mu_a)/(b + n_j)\) and \(\tilde{b} = b + n_j\). Note that throughout, \(\Sigma_j\) is as defined in models (3) - (8).

Let \(S^1_j = \sum_{i: z_i = j} (y_i - \tilde{y}_j)^T (y_i - \tilde{y}_j) + \frac{bn_j}{b + n_j} (\tilde{y}_j - \mu_a)^T (\tilde{y}_j - \mu_a)\) and \(S^2_j = \sum_{i: z_i = j} (y_i - \tilde{y}_j)^T + \frac{bn_j}{b + n_j} (\tilde{y}_j - \mu_a)(\tilde{y}_j - \mu_a)^T\). Also, let \([.]_r\) denote the \(r\)th diagonal element of the corresponding matrix. The full conditional distributions of the covariance
parameters are given as follows.

\[ M_1 : \quad \sigma^2 | y, z \sim \text{Inverse Gamma} \left( \tilde{s}/2, \tilde{S}/2 \right); \quad \tilde{s} = nd + s; \quad \tilde{S} = S + \sum_{j=1}^{K} S_j^1. \]

\[ M_2 : \quad \Sigma | y, z \sim \text{Inverse Wishart}(\tilde{s}, \tilde{S}); \quad \tilde{s} = s + n; \quad \tilde{S} = S + \sum_{j=1}^{K} S_j^2. \]

\[ M_3 : \quad \sigma^2 | y, z \sim \text{Inverse Gamma} \left( \tilde{s}/2, \tilde{S}/2 \right); \quad \tilde{s} = s + n; \quad \tilde{S} = \left[ SI + \sum_{j=1}^{K} S_j^2 \right]_r. \]

\[ M_4 : \quad \sigma^2 | y, z \sim \text{Inverse Gamma} \left( \tilde{s}/2, \tilde{S}/2 \right); \quad \tilde{s} = njd + s; \quad \tilde{S} = S + S_j^1. \]

\[ M_5 : \quad \sigma^2 | y, z \sim \text{Inverse Gamma} \left( \tilde{s}/2, \tilde{S}/2 \right); \quad \tilde{s} = s + nj; \quad \tilde{S} = \left[ S I + S_j^2 \right]_r. \]

\[ M_6 : \quad \Sigma | y, z \sim \text{Inverse Wishart}(\tilde{s}, \tilde{S}); \quad \tilde{s} = s + nj; \quad \tilde{S} = S + S_j^2. \] (a.6)

With these conditional posterior distributions, the Gibbs algorithm implemented is outlined as follows.

**Gibbs sampler**

1. Initialize the parameters \((\pi, \mu_j, \Sigma_j)\) and the hyperparameters (see (a.2)).

2. Update the allocation variables, \(z\): For \(i = 1, \ldots, n\), sample \(z_i\) from its conditional distribution given in (a.3).

3. Update \(\pi\): Sample \(\pi\) from its conditional distribution in (a.4).

4. Update \(\mu_j\): For each \(j \in \{1, \ldots, K\}\), draw \(\mu_j\) from the conditional distribution in (a.5)

5. Update \(\Sigma_j\): The full covariance matrices or the covariance parameters are drawn from their conditional distributions as given above.

6. Permute the component labels \(\{1, \ldots, K\}\) and align the parameters.

7. Repeat steps 2 - 6 for a desired number of iterations.

In the algorithm, the parameters were initialized using samples drawn from their prior distributions. In step 6, a random permutation of the \(K\) component labels as
proposed in Frühwirth-Schnatter (2001) is used to improve the mixing of the algorithm. We note that extensions of the Gibbs algorithm such as tempering MCMC (see, for example, Neal, 1996; Celeux et al., 2000) which have been proposed to handle trapping in local modes exist. However, we found that the Gibbs algorithm described above performed well in our application, producing a plausible clustering of the data and results comparable with other clustering approaches. All the algorithms for the proposed models have been coded in R (R Core Team, 2013).

A problem that occurs with Bayesian estimation of mixture models is the non-identifiability of the parameters which results from the invariance of the posterior distribution under a permutation of the parameters. Many solutions have been proposed for tackling this problem which include imposing identifiability constraints on the parameters and the use of relabelling algorithms (see Richardson and Green, 1997; Stephens, 2000). In this work, the relabelling algorithm of Stephens (2000) was applied retrospectively to the MCMC samples before these were used for further analysis.

To perform a Bayesian clustering of the HDSS sites using the finite Gaussian mixture model, we set $b = 1$ and $\delta_j = 1$ ($j = 1, \ldots, K$) in the prior distributions for $\mu_j$ and $\pi$, respectively, as these are natural choices. Also, the empirical mean of the data was used as the prior mean of $\mu_j$. In models (4) and (8), the prior distribution for $\Sigma$ and $\Sigma_j$ was Inverse Wishart($s, S = \text{diag}(1, \ldots, 1)$) while in models (3), (5), (6) and (7), we used Inverse Gamma($s/2, 1/2$) as prior for the covariance parameters $\sigma^2$, $\sigma^2_j$, $\sigma^2_{rj}$ and $\sigma^2_r$, respectively. We chose the values of $s$ using pilot runs in all the models, whilst noting that this hyperparameter must be greater than $d - 1$ in (4) and (8). Based on a previous analysis by Tatem et al. (2006), we considered $K = 2, \ldots, 10$. In all cases, the Gibbs algorithm was run for 20,000 iterations after a burn-in period of 20,000 iterations.
Table 1: INDEPTH HDSS sites in sub-Saharan Africa

<table>
<thead>
<tr>
<th>Site name</th>
<th>Country</th>
<th>Population</th>
<th>Catchment area (km²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Butajira</td>
<td>Ethiopia</td>
<td>74,400</td>
<td>262</td>
</tr>
<tr>
<td>2. Ifakara</td>
<td>Tanzania</td>
<td>161,000</td>
<td>2080</td>
</tr>
<tr>
<td>3. Rufiji</td>
<td>Tanzania</td>
<td>97,000</td>
<td>2313</td>
</tr>
<tr>
<td>4. Manhica</td>
<td>Mozambique</td>
<td>89,617</td>
<td>190</td>
</tr>
<tr>
<td>5. Agincourt</td>
<td>South Africa</td>
<td>87,040</td>
<td>97</td>
</tr>
<tr>
<td>6. Dikgale</td>
<td>South Africa</td>
<td>35,000</td>
<td>21</td>
</tr>
<tr>
<td>7. ACDIS</td>
<td>South Africa</td>
<td>93,500</td>
<td>983</td>
</tr>
<tr>
<td>8. Nouna</td>
<td>Burkina Faso</td>
<td>93,000</td>
<td>1841</td>
</tr>
<tr>
<td>9. Farafenni</td>
<td>Gambia</td>
<td>47,331</td>
<td>882</td>
</tr>
<tr>
<td>10. Navrongo</td>
<td>Ghana</td>
<td>156,735</td>
<td>2289</td>
</tr>
<tr>
<td>11. Bandim</td>
<td>Guinea Bissau</td>
<td>105,000</td>
<td>990</td>
</tr>
<tr>
<td>12. Bandafassi</td>
<td>Senegal</td>
<td>13,000</td>
<td>990</td>
</tr>
<tr>
<td>13. Mlomp</td>
<td>Senegal</td>
<td>8,200</td>
<td>80</td>
</tr>
<tr>
<td>14. Niakhar</td>
<td>Senegal</td>
<td>43,000</td>
<td>96</td>
</tr>
<tr>
<td>15. Rakai</td>
<td>Uganda</td>
<td>50,000</td>
<td>4581</td>
</tr>
<tr>
<td>16. Ouagadougou</td>
<td>Burkina Faso</td>
<td>82,387</td>
<td>182</td>
</tr>
<tr>
<td>17. Nairobi</td>
<td>Kenya</td>
<td>61,695</td>
<td>899</td>
</tr>
<tr>
<td>18. Kisumu</td>
<td>Kenya</td>
<td>230,000</td>
<td>619</td>
</tr>
<tr>
<td>19. Kintampo</td>
<td>Ghana</td>
<td>142,977</td>
<td>100</td>
</tr>
<tr>
<td>20. Karonga</td>
<td>Malawi</td>
<td>35,730</td>
<td>4087</td>
</tr>
<tr>
<td>21. Cross River</td>
<td>Nigeria</td>
<td>31,124</td>
<td>2057</td>
</tr>
<tr>
<td>22. Nahuche</td>
<td>Nigeria</td>
<td>136,106</td>
<td>902</td>
</tr>
<tr>
<td>23. Gilgel Gibe</td>
<td>Ethiopia</td>
<td>54,476</td>
<td>419</td>
</tr>
<tr>
<td>24. Magu</td>
<td>Tanzania</td>
<td>35,000</td>
<td>167</td>
</tr>
<tr>
<td>25. Kyamulibwa</td>
<td>Uganda</td>
<td>22,000</td>
<td>945</td>
</tr>
<tr>
<td>26. Iganga/Mayuge</td>
<td>Uganda</td>
<td>79,794</td>
<td>218</td>
</tr>
<tr>
<td>27. Chokwe</td>
<td>Mozambique</td>
<td>99,834</td>
<td>3137</td>
</tr>
<tr>
<td>28. MBITA</td>
<td>Kenya</td>
<td>54,014</td>
<td>187</td>
</tr>
<tr>
<td>29. Kombewa</td>
<td>Kenya</td>
<td>123,456</td>
<td>416</td>
</tr>
<tr>
<td>30. Kilifi</td>
<td>Kenya</td>
<td>260,000</td>
<td>1037</td>
</tr>
<tr>
<td>31. Dodowa</td>
<td>Ghana</td>
<td>111,976</td>
<td>1813</td>
</tr>
<tr>
<td>32. West Kiang</td>
<td>Gambia</td>
<td>14,364</td>
<td>803</td>
</tr>
<tr>
<td>33. Kilite Awlaelo</td>
<td>Ethiopia</td>
<td>65,848</td>
<td>768</td>
</tr>
<tr>
<td>34. Kersa</td>
<td>Ethiopia</td>
<td>52,480</td>
<td>174</td>
</tr>
<tr>
<td>35. Nanoro</td>
<td>Burkina Faso</td>
<td>61,632</td>
<td>649</td>
</tr>
<tr>
<td>36. Taabo</td>
<td>Côte d'Ivoire</td>
<td>38,478</td>
<td>629</td>
</tr>
<tr>
<td>37. Kaya</td>
<td>Burkina Faso</td>
<td>64,480</td>
<td>1090</td>
</tr>
<tr>
<td>38. Supone</td>
<td>Burkina Faso</td>
<td>86,089</td>
<td>803</td>
</tr>
<tr>
<td>39. Dabat</td>
<td>Ethiopia</td>
<td>46,984</td>
<td>379</td>
</tr>
</tbody>
</table>

Source: [www.indepth-network.org](http://www.indepth-network.org)
Table 2: Socioeconomic and environmental variables used in the study.

<table>
<thead>
<tr>
<th>Variable/Description</th>
<th>Period</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc50k travel time to nearest settlement of greater than 50,000 population (minutes)</td>
<td>2000&lt;sup&gt;1&lt;/sup&gt;</td>
</tr>
<tr>
<td>gecon measure of grid cell economic output in purchasing power parity(GCP by grid cell) = (population by grid cell) x (GCP/population)) by grid cell (USD)</td>
<td>2000&lt;sup&gt;2&lt;/sup&gt;</td>
</tr>
<tr>
<td>popdens population density (persons per km&lt;sup&gt;2&lt;/sup&gt;)</td>
<td>2010&lt;sup&gt;3&lt;/sup&gt;</td>
</tr>
<tr>
<td>alt altitude above sea level (metres)</td>
<td>1950-2000&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td>tmpmean annual mean temperature (°C x 10)</td>
<td>1950-2000&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td>pretot total precipitation (mm)</td>
<td>1950-2000&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td>tmpseas temperature seasonality (standard deviation x 100)(°C)</td>
<td>1950-2000&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
<tr>
<td>preseas precipitation seasonality (coefficient of variation)(mm)</td>
<td>1950-2000&lt;sup&gt;4&lt;/sup&gt;</td>
</tr>
</tbody>
</table>


Table 3: Summary statistics for the 39 HDSS sites

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min.</th>
<th>Ist Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>acc50k</td>
<td>0.0000</td>
<td>0.0995</td>
<td>0.1920</td>
<td>0.2778</td>
<td>0.3372</td>
<td>2.0320</td>
</tr>
<tr>
<td>log(gecon+0.5)</td>
<td>-1.1050</td>
<td>-0.5083</td>
<td>0.1661</td>
<td>0.1095</td>
<td>0.6510</td>
<td>2.0350</td>
</tr>
<tr>
<td>log(popdens)</td>
<td>-10.8900</td>
<td>2.1730</td>
<td>2.9860</td>
<td>2.7880</td>
<td>3.6750</td>
<td>8.6530</td>
</tr>
<tr>
<td>alt</td>
<td>0.0000</td>
<td>0.1761</td>
<td>0.6597</td>
<td>1.3930</td>
<td>2.6790</td>
<td>6.6220</td>
</tr>
<tr>
<td>tmpmean</td>
<td>3.1050</td>
<td>5.7980</td>
<td>6.8700</td>
<td>6.6160</td>
<td>7.4750</td>
<td>7.8590</td>
</tr>
<tr>
<td>pretot</td>
<td>0.9072</td>
<td>1.2740</td>
<td>1.5640</td>
<td>1.6700</td>
<td>1.9350</td>
<td>4.7720</td>
</tr>
<tr>
<td>tmpseas</td>
<td>0.1453</td>
<td>0.4966</td>
<td>0.7794</td>
<td>0.7228</td>
<td>1.0140</td>
<td>1.5630</td>
</tr>
<tr>
<td>preseas</td>
<td>0.6466</td>
<td>1.4670</td>
<td>2.0890</td>
<td>2.1220</td>
<td>2.8600</td>
<td>3.8050</td>
</tr>
</tbody>
</table>
Table 4: BIC values for the models

<table>
<thead>
<tr>
<th>Model</th>
<th>Number of components</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>2</td>
</tr>
<tr>
<td>$M_1$</td>
<td>755.44</td>
</tr>
<tr>
<td>$M_2$</td>
<td>700.07</td>
</tr>
<tr>
<td>$M_3$</td>
<td>673.75</td>
</tr>
<tr>
<td>$M_4$</td>
<td>755.12</td>
</tr>
<tr>
<td>$M_5$</td>
<td>690.24</td>
</tr>
<tr>
<td>$M_6$</td>
<td>732.30</td>
</tr>
</tbody>
</table>

Table 5: Clusters of the central clustering

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Sites</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Rakai, Kisumu, Magu, Kyamulibwa, Iganga/Mayuge, MBITA, Kombewa</td>
</tr>
<tr>
<td>2</td>
<td>Nouna, Farafenni, Navrongo, Bandim, Bandafassi, Mlomp, Niakhar, N’omoro, Kaya, Sapone, Ouagadougou, Cross River, West Kiang</td>
</tr>
<tr>
<td>3</td>
<td>Ifakara, Rufiji, Kintampo, Nahuche, Kilifi, Dodowa, Taabo</td>
</tr>
<tr>
<td>4</td>
<td>Butajira, Nairobi, Gilgel Gibe, Kilite Awlaelo, Kersa, Dabat</td>
</tr>
<tr>
<td>5</td>
<td>Manhica, Agincourt, Dikgale, ACDIS, Karonga, Chókwè</td>
</tr>
</tbody>
</table>
Figure 1: Plots of the socioeconomic and environmental layers. See Table 2 for details.
Figure 2: Plot of BIC against number of components for each of the six models.
Figure 3: A map of Africa showing the INDEPTH HDSS sites and their clusters. See Tables 1 and 5 for the numbering of the sites and their clusters, respectively.
Figure 4: Comparison of the central clustering with the $K$-means and hierarchical clustering. The clusterings are plotted in the space of the first two principal components of the data from the 39 HDSS sites.
Figure 5: Coverage map of the clusters. The clusters are numbered as given in Table 5. The uninhabited areas are coloured in grey.
Figure 6: Predictive probability map obtained from the $N = 18,496,629$ $1 \times 1$ km grid cells. Plotted is the maximum probability of being included in any cluster. Masked out, low population density areas, are coloured in grey. The existing 39 sites are shown as filled grey coloured circles. The 10 proposed new sites are shown in filled green coloured circles.
C. Edson Utazi,  
Southampton Statistical Sciences Research Institute 
University of Southampton 
Southampton, SO17 1BJ

Sujit K. Sahu,  
Southampton Statistical Sciences Research Institute 
University of Southampton 
Southampton, SO17 1BJ

Peter M. Atkinson,  
Faculty of Science and Technology, 
Lancaster University, 
Bailrigg, Lancaster LA1 4YW, UK

Natalia Tejedor  
Geography and Environment 
University of Southampton 
Southampton, SO17 1BJ

Andrew J. Tatem  
Geography and Environment 
University of Southampton 
Southampton, SO17 1BJ