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 1×1 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.[1] Moreover, new networks are being established, such as the Child Health and Mortality Prevention Surveillance Network (CHAMPS)[2], 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-

[1] See `http://www.indepth-network.org/index.php?option=com_content&task=view&id=1306&Itemid=1070` for details.

[2] See `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 1 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. For each site, the catchment area was digitised using the data and maps available on the INDEPTH website (`www.indepth-network.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[3] 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) tmpmean, annual mean temperature, (^0C), (iii) pretot, annual total precipitation, (iv) 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

[3] `http://forobs.jrc.ec.europa.eu/products/gam/`

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 2. 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 2) 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 1×1 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 1×1 km grid cells (see the last column of Table 1) which poses a problem of spatial misalignment while performing the comparison between each site and an arbitrary 1×1 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 1×1 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.

3 Bayesian model-based clustering methods

3.1 The Bayesian finite Gaussian mixture model

Let $\mathbf{y}_i \in \mathbb{R}^d$ denote the d average measurements from the ith HDSS site, $i = 1, \ldots, n = 39$. To group these n d-dimensional observations into an unknown number of clusters, we assume the following Gaussian mixture model with K components:

$$\mathbf{y}_i|\boldsymbol{\Theta}_K \sim \sum_{j=1}^{K} \pi_j \mathcal{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_j), \quad i = 1, \ldots, n; \tag{1}$$

where $\boldsymbol{\Theta}_K = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$, $\boldsymbol{\theta}_j = (\pi_j, \boldsymbol{\mu}_j, \boldsymbol{\Sigma}_j)$, and the parameters $\boldsymbol{\mu}_j$, $\boldsymbol{\Sigma}_j$ and π_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 jth 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 ith site. That is, $z_i = j$ if the ith site belongs to the jth cluster ($j = 1, \ldots, K$). For future reference, let $\mathbf{z} = (z_1, \ldots, z_n)$ denote the unknown cluster indicators and $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$ denote the full set of data from the 39 HDSS sites. The model can now be expressed in a hierarchical form as

$$\begin{aligned} &\mathbf{y}_i|z_i = j, \boldsymbol{\Theta}_K \sim \mathcal{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_j); \\ &p(z_i = j|\boldsymbol{\Theta}_K) = \pi_j. \end{aligned} \tag{2}$$

In the model, each component is characterized by its mean vector $\boldsymbol{\mu}_j$ and covariance matrix $\boldsymbol{\Sigma}_j$. The $\boldsymbol{\mu}_j$'s form the centers of the clusters whereas the $\boldsymbol{\Sigma}_j$'s determine the geometric features of the clusters; see Fraley and Raftery (2007). Each $\boldsymbol{\Sigma}_j$ 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:

$$\begin{aligned}
\mathcal{M}_1: &\quad \boldsymbol{\Sigma}_j = \sigma^2 \boldsymbol{I} \quad \forall \quad j; && (3)\\
\mathcal{M}_2: &\quad \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma} \quad \forall \quad j; && (4)\\
\mathcal{M}_3: &\quad \boldsymbol{\Sigma}_j = \text{diag}(\sigma_1^2, \ldots, \sigma_d^2) \quad \forall \quad j; && (5)\\
\mathcal{M}_4: &\quad \boldsymbol{\Sigma}_j = \sigma_j^2 \boldsymbol{I}; && (6)\\
\mathcal{M}_5: &\quad \boldsymbol{\Sigma}_j = \text{diag}(\sigma_{1j}^2, \ldots, \sigma_{dj}^2); && (7)\\
\mathcal{M}_6: &\quad \boldsymbol{\Sigma}_j = \boldsymbol{\Sigma}_j; && (8)
\end{aligned}$$

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 $\boldsymbol{\Sigma}_j$ 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

$$BIC = -2 \log p(\hat{\boldsymbol{\Theta}}_{\mathcal{M},K}; \mathbf{y}) + \gamma_K \log n, \tag{9}$$

where $p(.)$ is the 'complete-data' likelihood of model $\mathcal{M}$ with K components evaluated using the maximum a posteriori (MAP) estimate, n is the sample size and γ_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 $\hat{\boldsymbol{\Theta}}_{\mathcal{M},K}$ to be the MAP estimate given by:

$$\hat{\boldsymbol{\Theta}}_{\mathcal{M},K} \approx \arg \max_{\boldsymbol{\Theta}^{(t)}_{\mathcal{M},K}; 1 \leq t \leq T} \left\{ p(\boldsymbol{\Theta}^{(t)}_{\mathcal{M},K}; \mathbf{y}) \times p(\boldsymbol{\Theta}^{(t)}_{\mathcal{M},K}) \right\}, \tag{10}$$

where $p(\boldsymbol{\Theta}_{\mathcal{M},K})$ denotes the prior distribution for $\boldsymbol{\Theta}_{\mathcal{M},K}$. 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 configuration 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, $\mathbf{z}_1, \ldots, \mathbf{z}_T$, 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, $\mathbf{z}^*$ is a central clustering if for a given small $\epsilon > 0$ (typically, $0 < \epsilon < 1$),

$$\mathbf{z}^* = \arg \max_{1 \leq t \leq T} \frac{1}{T} \# \left\{ \mathbf{z}_l; 1 \leq l \leq T : d(\mathbf{z}_t, \mathbf{z}_l) < \epsilon \right\}, \tag{11}$$

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 ϵ 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 ith cluster of $\mathbf{z}_l$ and jth cluster of $\mathbf{z}_t$ 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 $\mathbf{z}_l$ and $\mathbf{z}_t$ is given by

$$d(\mathbf{z}_l, \mathbf{z}_t) = \max \left\{ \tilde{d}(\mathbf{z}_l, \mathbf{z}_t), \tilde{d}(\mathbf{z}_t, \mathbf{z}_l) \right\}, \tag{12}$$

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

$$\frac{1}{T} \# \{\mathbf{z}_t; 1 \leq t \leq T : d(\mathbf{z}_t, \mathbf{z}^*) < \epsilon^*\} \approx 0.95. \tag{13}$$

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 1×1 km grid cells. Let $\mathbf{y}_m$ denote the 8-dimensional socioeconomic and environmental measurements for the mth 1×1 km grid cell, $m = 1, \ldots, N(= 18,496,629)$. Let z_m denote the unknown cluster index for $\mathbf{y}_m$. Our objective here is to find, $p(z_m = j | \mathbf{y}, \mathbf{y}_m, k^*)$ for $j = 1, \ldots, k^*$ where k^* is the chosen value of the optimum number of clusters. This probability can be expressed as

$$p(z_m = j | \mathbf{y}, \mathbf{y}_m, k^*) = \int p(z_m = j | \mathbf{y}_m, \mathbf{\Theta}_{k^*}, \mathbf{z}) p(\mathbf{\Theta}_{k^*}, \mathbf{z} | \mathbf{y}, \mathbf{y}_m) d\mathbf{\Theta}_{k^*} d\mathbf{z}, \tag{14}$$

where $\boldsymbol{\Theta}_{k*}$ 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 $\mathbf{y}_m$'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:

$$\begin{aligned} p(z_m = j|\mathbf{y}, \mathbf{y}_m, k^*) &\approx \int p(z_m = j|\mathbf{y}_m, \boldsymbol{\Theta}_{k^*}, \mathbf{z}) p(\boldsymbol{\Theta}_{k^*}, \mathbf{z}|\mathbf{y}) d\boldsymbol{\Theta}_{k^*} d\mathbf{z}, \\ &= E_{\boldsymbol{\Theta}_{k^*}, \mathbf{z}} \left(p(z_m = j|\mathbf{y}_m, \boldsymbol{\Theta}_{k^*}, \mathbf{z}) \right). \end{aligned} \tag{15}$$

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

$$p(z_m = j|\mathbf{y}_m, \boldsymbol{\Theta}_{k^*}, \mathbf{z}) = \frac{\pi_j \mathcal{N}_d(\mathbf{y}_m; \boldsymbol{\theta}_j)}{\sum_{l=1}^{k^*} \pi_l \mathcal{N}_d(\mathbf{y}_m; \boldsymbol{\theta}_l)}, \quad j = 1, \ldots, k^*, \tag{16}$$

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

$$\hat{z}_m = \underset{j}{\operatorname{argmax}} \{ p(z_m = j|\mathbf{y}, \mathbf{y}_m, k^*) \}. \tag{17}$$

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 $N_0 = \#\{m : \max\{p(z_m = j|\mathbf{y}, \mathbf{y}_m, k^*)\} < p_0; 1 \leq m \leq N, 1 \leq j \leq k^*\}$ denote the number of grid cells whose classification probabilities were less than the threshold probability, p_0. Let $D = \{\mathbf{s}_1, \ldots, \mathbf{s}_{N_0}\}$ 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

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

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 socioeconomic 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 $\mathcal{M}_1$ to $\mathcal{M}_6$ for the given values of K. The BIC values are also plotted in Figure 3. It can be seen that the constrained models, $\mathcal{M}_1$ to $\mathcal{M}_5$, are clearly better than the full model, $\mathcal{M}_6$ according to the BIC. Further, models $\mathcal{M}_2$ and $\mathcal{M}_3$ with the same covariance matrix for all the components fit the data better than other constrained models. The best fitting model is $\mathcal{M}_3$ (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 ($\mathbf{z}^*$) provided in Table 5. The value of ϵ corresponding to $\mathbf{z}^*$ is 0.05. (This ϵ value was determined as the center of the distances between the clusterings. We note that for other values within ± 0.03 of this ϵ value and for $\epsilon \to 0$, no other central clustering was detected.) Figure 4 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 $\{\mathbf{z} : d(\mathbf{z}^*, \mathbf{z}) < 0.1282\}$. 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 5. (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 1×1 km grid cells across Africa as discussed in Section 4.1. The results of this analysis are displayed in Figures 6 and 7. From the cluster map in Figure 6, 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 *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 2c - 2f. Lastly, the sites in Cluster 5 typically characterize the steppes and semi-desert regions of the continent.

The probability map in Figure 7 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 ≈ 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 7. 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 1×1 km spatial resolution. The central clustering of the HDSS sites, the cluster and probability maps resulting from predictive clustering of the 1×1 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 1×1 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.

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. 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

Baiden, F., Hodgson, A., and Binka, F. N. (2006). Demographic surveillance sites and emerging challenges in international health. Bull. World Healhth Organ., **84**, 163.

Banfield, J. D. and Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. Biometrics, **49**, 803–821.

Bouveyron, C. and Brunet-Saumard, C. (2014). Model-based clustering of high-dimensional data: A review. Computational Statistics and Data Analysis, **71**, 52–78.

Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition, **28**, 781–793.

Celeux, G., Hurn, M., and Robert, C. P. (2000). Computational and inferential dif-

ficulties with mixture posterior distributions. Journal of the American Statistical Association, **95**(451), 957–970.

Celeux, G., Forbes, F., Robert, C. P., and Titterington, D. M. (2006). Deviance information criteria for missing data models. Bayesian Analysis, **1**(4), 651–673.

Chipeta, M. G., Terlouw, D. J., Phiri, K. S., and Diggle, P. J. (2016). Adaptive geostatistical design and analysis for prevalence surveys. Spatial Statistics, page doi:10.1016/j.spasta.2015.12.004.

Dash, J., Jeganathan, C., and Atkinson, P. M. (2010). The use of MERIS terrestrial chlorophyll index to study spatio-temporal variation in vegetation phenology over India. Remote Sensing of Environment, **114**(7), 1388–1402.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B, **39**(1), 1–38.

Dey, D. K., Kuo, L., and Sahu, S. K. (1995). A Bayesian predictive approach to determining the number of components in a mixture distribution. Statistics and Computing, **5**, 297–305.

Diebolt, J. and Robert, C. P. (1994). Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society, Series B, **56**, 363–375.

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association, **90**(430), 577–588.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis and density estimation. Journal of the American Statistical Association, **97**, 611–631.

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for Normal mixture estimaton and model-based clustering. Journal of Classification, **24**, 155–181.

Frühwirth-Schnatter, S. (2001). Markov chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association, **96**(453), 194–209.

Hijmans, R. J., Cameron, S. E., Parra, J. L., Jones, P. G., and Jarvis, A. (2005). Very high resolution interpolated climate surfaces for global land areas. International Journal of Climatology, **25**, 1965–1978.

Jia, P., Sankoh, O., and Tatem, A. J. (2015). Mapping the environmental and socioeconomic coverage of the INDEPTH international health and demographic surveillance system network. Health & Place, **36**, 88–96.

Kass, R. E. and Raftery, A. E. (1995). Bayes factors. Journal of the American Statistical Association, **90**(430), 773–795.

Linard, C., Gilbert, M., Snow, R. W., Noor, A. M., and Tatem, A. J. (2012). Population distribution, settlement patterns and accessibility across Africa in 2010. PLoS ONE, **7**(2).

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In: Cam, L. M., Neyman, J. (Eds.), Proceedngs of the 5th Berkeley Symposium on Mathematical Sciences and Probability, Vol. 1. University of California Press, pages 281–297.

McLachlan, G. J. and Peel, D. (2000). *Finite mixture models*. Wiley Interscience, New York.

Mukhopadhyay, S., Bhattacharya, S., and Dihidar, K. (2011). On Bayesian "central clustering": Application to landscape classification of Western Ghats. The Annals of Applied Statstics, **5**(3), 1948–1977.

Neal, R. M. (1996). Sampling from multimodal distributions using tempered transitions. Statsitics and Computing, **6**, 353–366.

Ngom, P., Binka, F. N., Phillips, J. F., Pence, B., and Macloed, B. (2001). Demographic surveillance and health equity in sub-Saharan Africa. Health Policy and Planning, **16**, 337–344.

Nordhaus, W. D. (2006). Geography and macroeconomics: New data and new findings. Proceedings of the National Academy of Science, **103**(10), 3510–17.

R Core Team (2013). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Richardson, S. and Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components. J. R. Statist. Soc. B, **59**(4), 731–792.

Royle, J. A. and Nychka, D. (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. Computers & Geosciences, **24**(5), 497–488.

Sahu, S. K. and Cheng, R. C. H. (2003). A fast distance based approach for determining the number of components in mixtures. The Canadian Journal of Statistics, **31**, 3–22.

Sankoh, O. and Byass, P. (2012). The INDEPTH network: Filling vital gaps in global epidemiology. International Journal of Epidemiology, **41**, 579–588.

Schwarz, G. (1978). Estimating the dimension of a model. Annals of Statistics, **6**, 461–464.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and van der Linde, A. (2002). Bayesian measures of model complexity and fit. Journal of the Royal Statistical Society, Series B, **64**, 583–616.

Steele, R. J. and Raftery, A. E. (2009). Performance of Bayesian model selection criteria for Gaussian mixture models. Dept. Stat., Univ. Washington, Wasington, DC, Tech. Rep. 559.

Stepehens, M. (2000). Dealing with label switching in mixture models. Journal of the Royal Statistical Society, Series B, **62**(4), 795–809.

Tatem, A. J., Snow, R. W., and Hay, S. I. (2006). Mapping the environmental coverage of the INDEPTH demographic surveillance system network in rural Africa. Trop. Med. Int. Health, **11**(8), 1318–1326.

Ward, J. H. (1963). Hierarchical groupings to optimize an objective function. Journal of the American Statistical Association, **105**, 234–244.

WHO (2013). Strengthening civil registration and vital statistics for births, deaths and causes of death: Resource kit. *Luxembourg:* World Health Organisation.

Ye, Y., Wamukoya, M., Ezeh, A., Emina, J. B. O., and Sankoh, O. (2012). Health and demographic surveillance systems: A step towards full civil registration and vital statistics system in sub-Saharan Africa? BMC Public Health, **12**(741).

Appendix: posterior details and the Gibbs sampler

Given the set of the measurements of the HDSS sites and their cluster indicators $\{(\mathbf{y}_i, z_i); i = 1, \ldots, n\}$, the 'complete-data' likelihood of the model is

$$p(\boldsymbol{\Theta}_K; \mathbf{y}, \mathbf{z}) = \prod_{j=1}^{K} \prod_{i:z_i=j} \pi_j \mathcal{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_j), \tag{a.1}$$

from which the observed-data likelihood can be easily obtained as $p(\boldsymbol{\Theta}_K; \mathbf{y}) = \int p(\boldsymbol{\Theta}_K; \mathbf{y}, \mathbf{z}) d\mathbf{z}$ 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 ($j = 1, \ldots, K; r = 1, \ldots, d$):

$$\begin{aligned} \pi_1, \ldots, \pi_K &\sim \text{Dir}(\delta_1, \ldots, \delta_K); \\ \boldsymbol{\mu}_j | \boldsymbol{\Sigma}_j &\sim \mathcal{N}_d(\boldsymbol{\mu}_a, 1/b\boldsymbol{\Sigma}_j); \\ \boldsymbol{\Sigma}_j, \boldsymbol{\Sigma} &\sim \text{Inverse Wishart}(s, \boldsymbol{S}); \\ \sigma^2, \sigma_j^2, \sigma_{rj}^2, \sigma_r^2 &\sim \text{Inverse Gamma}(s/2, S/2). \end{aligned} \tag{a.2}$$

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 $p(\boldsymbol{\Theta}_K; \mathbf{y}, \mathbf{z}) \times p(\boldsymbol{\Theta}_K)$, which is simply the product of the likelihood in (a.1) and the joint prior distribution on the parameters, $p(\boldsymbol{\Theta}_K)$. 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

$$p(z_i = j | \mathbf{y}, \boldsymbol{\Theta}_K, \mathbf{z}_{-i}) = \frac{\pi_j \mathcal{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_j)}{\sum_{l=1}^{K} \pi_l \mathcal{N}_d(\mathbf{y}_i; \boldsymbol{\theta}_l)}; \; j = 1, \ldots, K. \tag{a.3}$$

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

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

For the component means, $\boldsymbol{\mu}_j$, we have

$$\boldsymbol{\mu}_j|\mathbf{y},\mathbf{z},\boldsymbol{\Sigma}_j \sim \mathcal{N}_d(\tilde{\boldsymbol{\mu}}_a, 1/\tilde{b}\boldsymbol{\Sigma}_j); \tag{a.5}$$

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

Let $S_j^1 = \sum_{i:z_i=j}(\mathbf{y}_i - \bar{\mathbf{y}}_j)^T(\mathbf{y}_i - \bar{\mathbf{y}}_j) + \frac{bn_j}{b+n_j}(\bar{\mathbf{y}}_j - \boldsymbol{\mu}_a)^T(\bar{\mathbf{y}}_j - \boldsymbol{\mu}_a)$ and $\boldsymbol{S}_j^2 = \sum_{i:z_i=j}(\mathbf{y}_i - \bar{\mathbf{y}}_j)(\mathbf{y}_i - \bar{\mathbf{y}}_j)^T + \frac{bn_j}{b+n_j}(\bar{\mathbf{y}}_j - \boldsymbol{\mu}_a)(\bar{\mathbf{y}}_j - \boldsymbol{\mu}_a)^T$. Also, let $[.]_r$ denote the rth diagonal element of the corresponding matrix. The full conditional distributions of the covariance parameters are given as follows.

$$\begin{aligned}
\mathcal{M}_1: &\quad \sigma^2|\mathbf{y},\mathbf{z} \sim \text{Inverse Gamma}\left(\tilde{s}/2, \tilde{S}/2\right); \quad \tilde{s} = nd + s; \quad \tilde{S} = S + \sum_{j=1}^{K} S_j^1.\\
\mathcal{M}_2: &\quad \boldsymbol{\Sigma}|\mathbf{y},\mathbf{z} \sim \text{Inverse Wishart}(\tilde{s}, \tilde{\boldsymbol{S}}); \quad \tilde{s} = s + n; \quad \tilde{\boldsymbol{S}} = \boldsymbol{S} + \sum_{j=1}^{K} \boldsymbol{S}_j^2.\\
\mathcal{M}_3: &\quad \sigma_r^2|\mathbf{y},\mathbf{z} \sim \text{Inverse Gamma}\left(\tilde{s}/2, \tilde{S}/2\right); \quad \tilde{s} = s + n; \quad \tilde{S} = \left[S\boldsymbol{I} + \sum_{j=1}^{K} \boldsymbol{S}_j^2\right]_r.\\
\mathcal{M}_4: &\quad \sigma_j^2|\mathbf{y},\mathbf{z} \sim \text{Inverse Gamma}\left(\tilde{s}/2, \tilde{S}/2\right); \quad \tilde{s} = n_j d + s; \quad \tilde{S} = S + S_j^1.\\
\mathcal{M}_5: &\quad \sigma_{rj}^2|\mathbf{y},\mathbf{z} \sim \text{Inverse Gamma}\left(\tilde{s}/2, \tilde{S}/2\right); \quad \tilde{s} = s + n_j; \quad \tilde{S} = \left[S\boldsymbol{I} + \boldsymbol{S}_j^2\right]_r.\\
\mathcal{M}_6: &\quad \boldsymbol{\Sigma}_j|\mathbf{y},\mathbf{z} \sim \text{Inverse Wishart}(\tilde{s}, \tilde{\boldsymbol{S}}); \quad \tilde{s} = s + n_j; \quad \tilde{\boldsymbol{S}} = \boldsymbol{S} + \boldsymbol{S}_j^2.
\end{aligned} \tag{a.6}$$

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

Gibbs sampler

1. Initialize the parameters ($\boldsymbol{\pi}$, $\boldsymbol{\mu}_j$ and $\boldsymbol{\Sigma}_j$) and the hyperparameters (see (a.2)).
2. Update the allocation variables, $\mathbf{z}$: For $i = 1, \ldots, n$, sample z_i from its conditional distribution given in (a.3).
3. Update $\boldsymbol{\pi}$: Sample $\boldsymbol{\pi}$ from its conditional distribution in (a.4).
4. Update $\boldsymbol{\mu}_j$: For each $j \in \{1, \ldots, K\}$, draw $\boldsymbol{\mu}_j$ from the conditional distribution in (a.5)

5. Update $\boldsymbol{\Sigma}_j$: 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 of 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; Stepehens, 2000). In this work, the relabelling algorithm of Stepehens (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 $\boldsymbol{\mu}_j$ and $\boldsymbol{\pi}$, respectively, as these are natural choices. Also, the empirical mean of the data was used as the prior mean of $\boldsymbol{\mu}_j$. In models (4) and (8), the prior distribution for $\boldsymbol{\Sigma}$ and $\boldsymbol{\Sigma}_j$ was Inverse Wishart$(s, \boldsymbol{S} = \text{diag}(1, \ldots, 1))$ while in models (3), (5), (6) and (7), we used Inverse Gamma$(s/2, 1/2)$ as prior for the covariance parameters σ^2, σ_j^2, σ_{rj}^2 and σ_r^2, 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

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

Source: `www.indepth-network.org`

Table 2: Socioeconomic and environmental variables used in the study.

Variable/ Layer	Description	Period
• acc50k	travel time to nearest settlement of greater than 50,000 population (minutes)	2000[1]
• gecon	measure of grid cell economic output in purchasing power parity(GCP by grid cell =(population by grid cell) × (GCP/ population)) by grid cell (USD)	2000[2]
• popdens	population density (persons per km^2)	2010[3]
• alt	altitude above sea level (metres)	1950-2000[4]
• tmpmean	annual mean temperature (°C × 10)	1950-2000[4]
• pretot	total precipitation (mm)	1950-2000[4]
• tmpseas	temperature seasonality (standard deviation × 100)(°C)	1950-2000[4]
• preseas	precipitation seasonality (coefficient of variation)(mm)	1950-2000[4]

Source: [1]`http://forobs.jrc.ec.europa.eu/products/gam/`; [2]`http://gecon.yale.edu/`; [3]`www.worldpop.org.uk`; [4]`http://www.worldclim.org/`.

Table 3: Summary statistics for the 39 HDSS sites

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

Table 4: BIC values for the models

	Number of components								
Model	2	3	4	5	6	7	8	9	10
$\mathcal{M}_1$	755.44	736.88	754.14	752.73	757.65	780.15	797.35	862.73	886.49
$\mathcal{M}_2$	700.07	706.47	691.51	677.41	678.71	705.58	735.57	757.21	783.38
$\mathcal{M}_3$	673.75	655.03	654.13	643.51	653.76	667.40	674.31	717.55	759.38
$\mathcal{M}_4$	755.12	725.43	733.03	729.54	775.42	782.45	820.11	865.03	891.06
$\mathcal{M}_5$	690.24	703.45	762.03	763.87	835.50	979.84	991.90	1033.41	1152.67
$\mathcal{M}_6$	732.30	870.14	1056.34	1160.66	1360.93	1524.65	1661.94	1862.36	2040.59

Table 5: Clusters of the central clustering

Cluster	Sites
1	Rakai, Kisumu, Magu, Kyamulibwa, Iganga/Mayuge, MBITA, Kombewa
2	Nouna, Farafenni, Navrongo, Bandim, Bandafassi, Mlomp, Niakhar, Nanoro, Kaya, Sapone, Ouagadougou, Cross River, West Kiang
3	Ifakara, Rufiji, Kintampo, Nahuche, Kilifi, Dodowa, Taabo
4	Butajira, Nairobi, Gilgel Gibe, Kilite Awlaelo, Kersa, Dabat
5	Manhica, Agincourt, Dikgale, ACDIS, Karonga, Chókwè

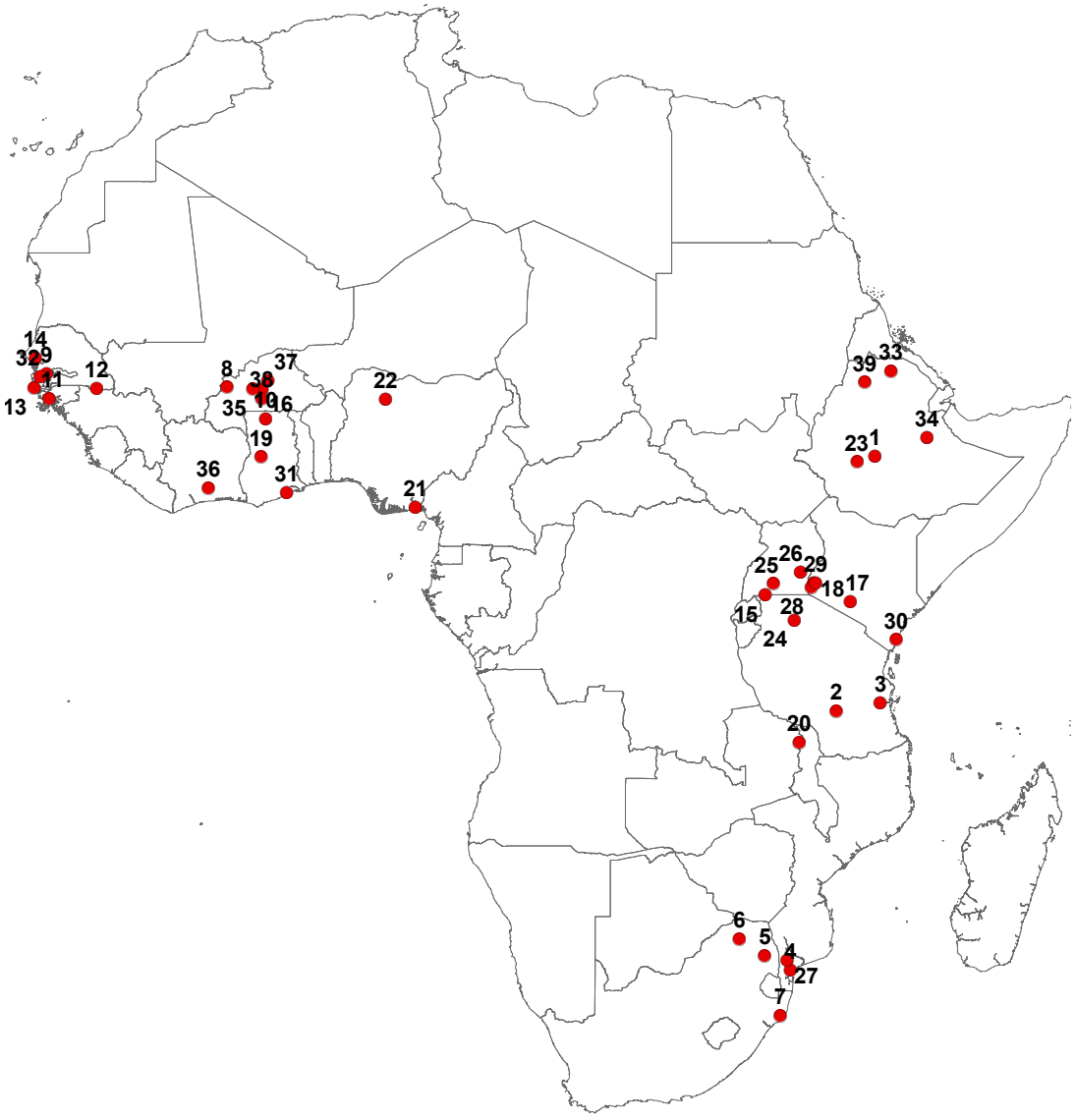

Figure 1: A map of Africa showing the locations of the INDEPTH HDSS sites. The sites are numbered as given in Table 1.

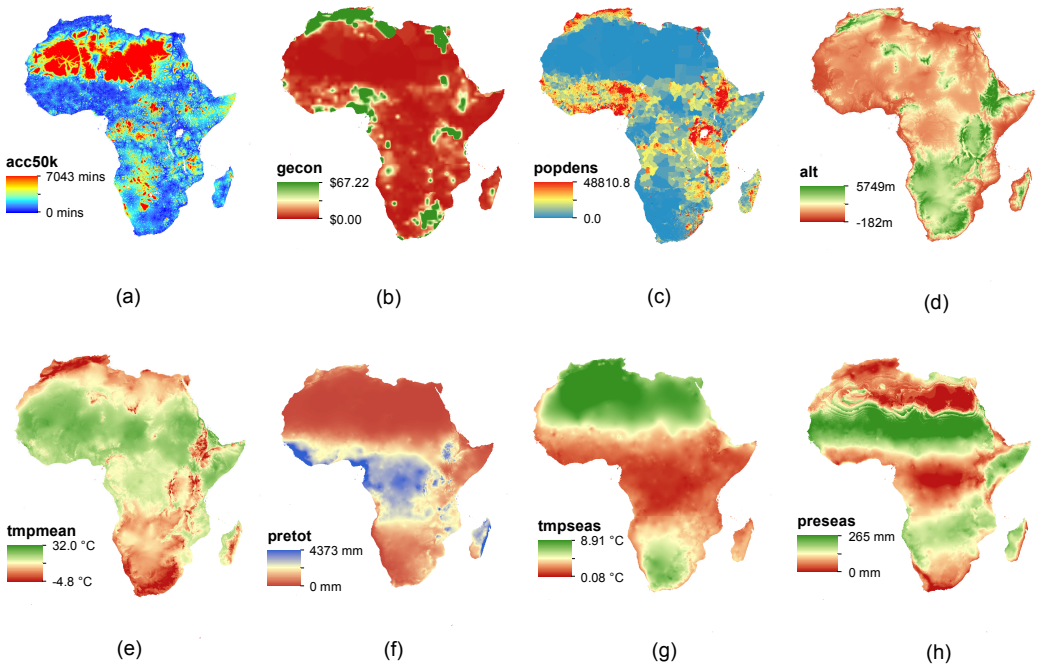

Figure 2: Plots of the socioeconomic and environmental layers. See Table 2 for details.

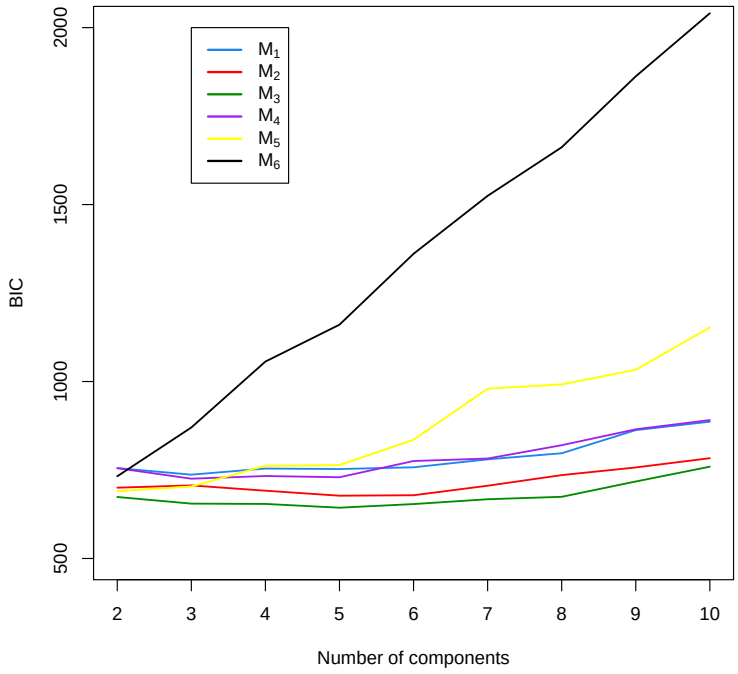

Figure 3: Plot of BIC against number of components for each of the six models.

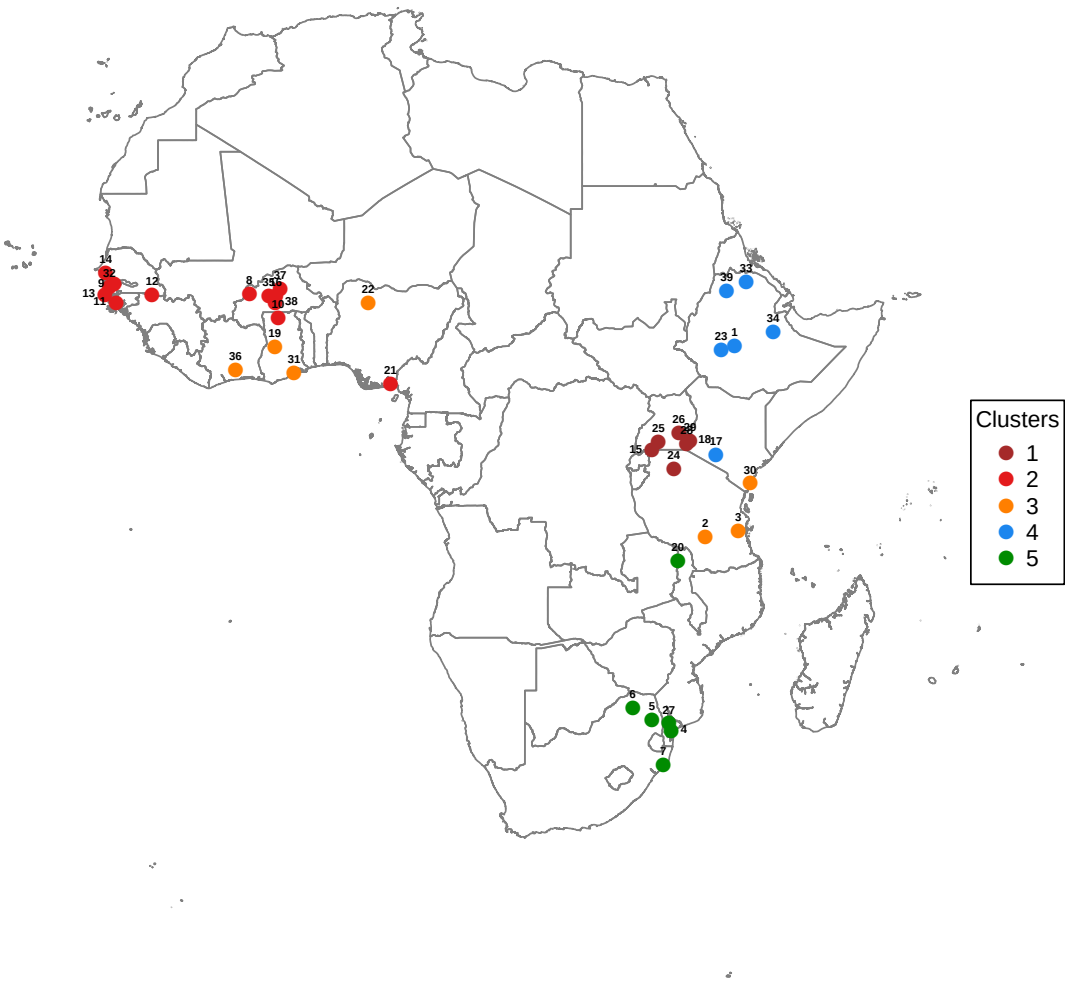

Figure 4: Map of Africa showing the sites and their clusters. See Tables 1 and 5 for the numbering of the sites and their clusters, respectively.

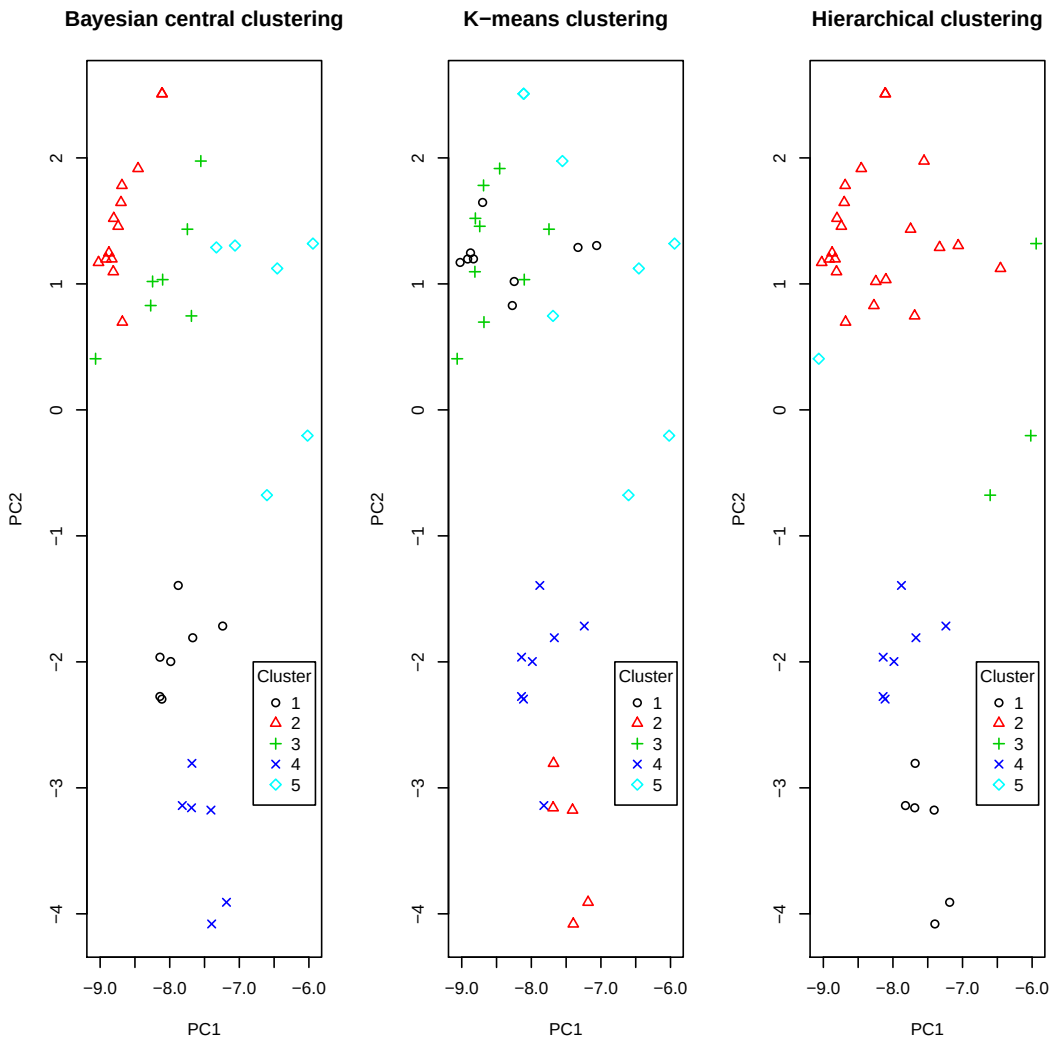

Figure 5: Comparison of the central clustering with the K-means and hierarchical clustering. The clusterings are plotted in the space of the first two principal components of the data from the 39 HDSS sites.

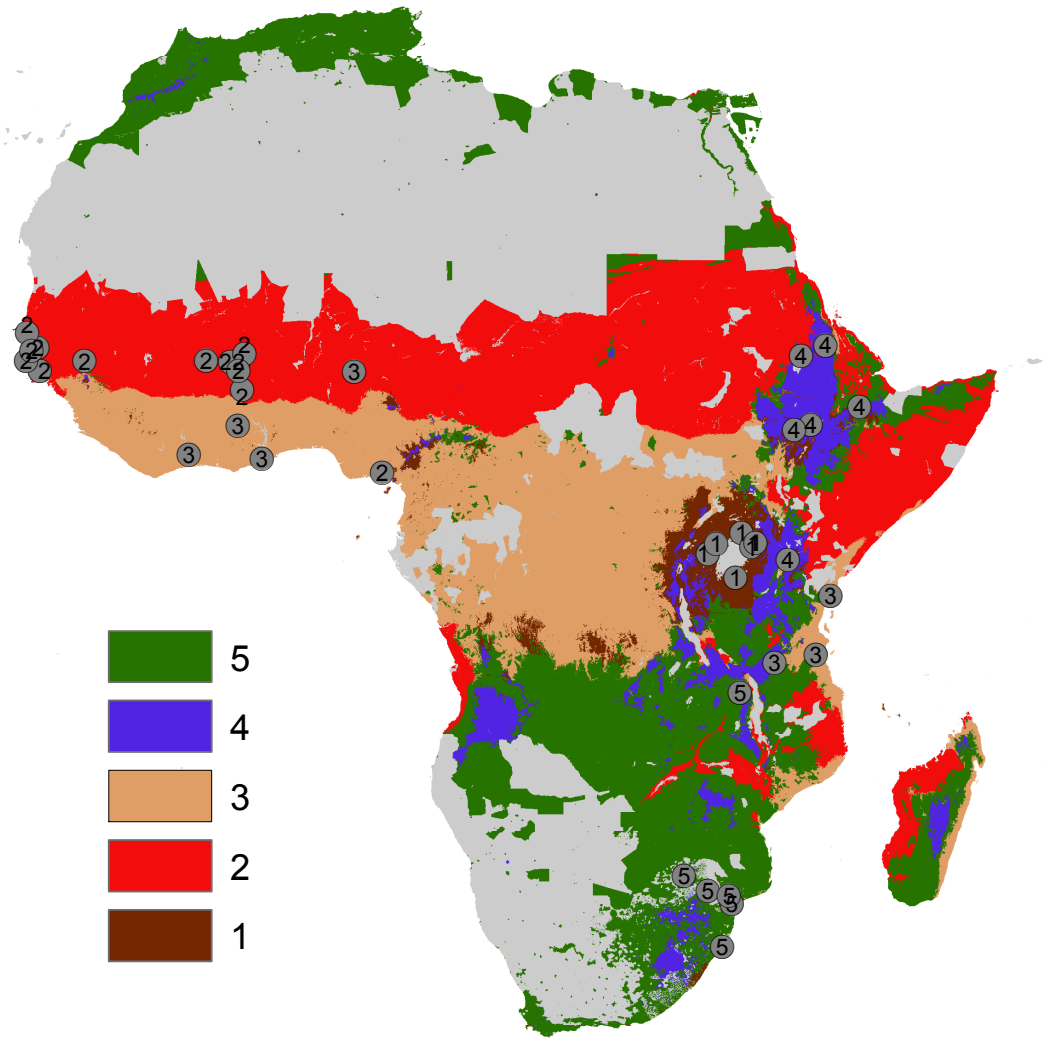

Figure 6: Coverage map of the clusters. The clusters are numbered as given in Table 5. The uninhabited areas are coloured in grey.

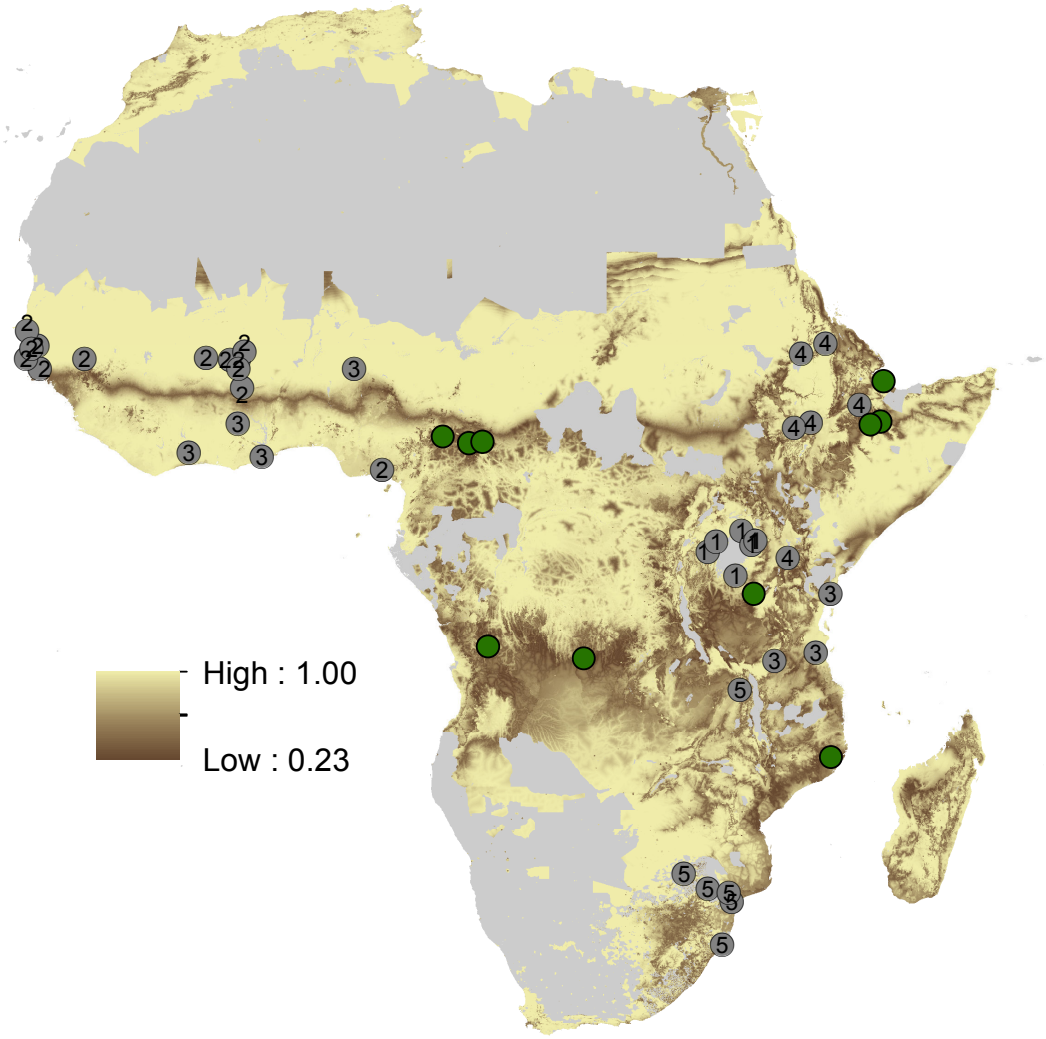

Figure 7: Predictive probability map obtained from the $N = 18,496,629$ 1×1 km grid cells. Plotted is the maximum probability of being included in any cluster. Masked out, low population density areas, are coloured in grey. The existing 39 sites are shown as filled grey coloured circles. The 10 proposed new sites are shown in filled green coloured circles.