Random forest variable selection in spatial malaria transmission modelling in Mpumalanga Province, South Africa

Thandi Kapwata,[1,2] Michael T. Gebreslasie[2]

[1]Biostatistics Unit, South African Medical Research Council, Cape Town; [2]School of Agriculture, Earth, and Environmental Sciences, University of KwaZulu-Natal, Durban, South Africa

Abstract

Malaria is an environmentally driven disease. In order to quantify the spatial variability of malaria transmission, it is imperative to understand the interactions between environmental variables and malaria epidemiology at a micro-geographic level using a novel statistical approach. The random forest (RF) statistical learning method, a relatively new variable-importance ranking method, measures the variable importance of potentially influential parameters through the percent increase of the mean squared error. As this value increases, so does the relative importance of the associated variable. The principal aim of this study was to create predictive malaria maps generated using the selected variables based on the RF algorithm in the Ehlanzeni District of Mpumalanga Province, South Africa. From the seven environmental variables used [temperature, lag temperature, rainfall, lag rainfall, humidity, altitude, and the normalized difference vegetation index (NDVI)], altitude was identified as the most influential predictor variable due its high selection frequency. It was selected as the top predictor for 4 out of 12 months of the year, followed by NDVI, temperature and lag rainfall, which were each selected twice. The combination of climatic variables that produced the highest prediction accuracy was altitude, NDVI, and temperature. This suggests that these three variables have high predictive capabilities in relation to malaria transmission. Furthermore, it is anticipated that the predictive maps generated from predictions made by the RF algorithm could be used to monitor the progression of malaria and assist in intervention and prevention efforts with respect to malaria.

Correspondence: Thandi Kapwata, Biostatistics Unit, South African Medical Research Council, Francie van Zijl Drive, PO Box 19070, 7505 Cape Town, South Africa.
Tel: +27.21.938.0911 - Fax: +27.21.938.0200.
E-mail: thandi.kapwata@mrc.ac.za

Key words: Random forest; Modelling malaria transmission; South Africa.

Received for publication: 1 December 2015.
Revision received: 27 April 2016.
Accepted for publication: 15 May 2016.

Geospatial Health 2016; 11:434
doi:10.4081/gh.2016.434

Introduction

Malaria transmission in South Africa has been decreasing over the past years and is now limited to the low-lying north-eastern parts of Limpopo, Mpumalanga and KwaZulu-Natal Provinces (Gerritsen *et al.*, 2008). Mpumalanga Province has maintained a successful control programme encompassing rapid detection and treatment of confirmed malaria cases at primary health care facilities and vector control through indoor residual spraying with insecticides and focal larviciding (Govere *et al.*, 2000). However, the province still contributes to 44% of the country's notified malaria cases (Silal *et al.*, 2013).

Climate has a great impact on the life cycle of the mosquito vector and malaria parasite (Craig *et al.*, 1999) and studies have shown that malaria risk and transmission intensity exhibit significant spatial and temporal variability related to variations in climate, altitude, topography and socio-economic activities (Brooker *et al*., 2004; Gosoniu *et al*., 2006; Castillo Riquelme *et al*., 2008; Karthe, 2010). Therefore, there is an urgent need to be able to identify which climatic variables have the greatest influence on malaria transmission as this allows the modelling, both spatially and temporally, of malaria transmission under various climatic conditions. An interpretation of the relationship and interactions among the most significant environmental variables and malaria at a more detailed level is also an important part in developing malaria early warning systems (EWS), identifying potential outbreaks, and targeting vector control strategies (Ngomane and De Jager, 2012).

Statistical methods add an important dimension to prediction models. However, for statistical methods to be applicable for selecting significant variables to model malaria transmission patterns in space and time, thus resulting in efficient targeting of interventions against the disease, accurate disease and climatic data are required (Ostfeld *et al.*, 2005). Variable selection is a procedure that is important in the process of creating such prediction models, therefore parsimony (selecting few strong predictors that are easily interpretable from many potential candidates) and plausibility (association with malaria being etiologically explainable) are key.

Different analytical approaches towards variable selection of varying sophistication have been employed, but these approaches are complex and require a detailed knowledge of statistical methods. Kleinshmidt *et al.* (2000), Craig *et al.* (2007) and Mabaso *et al*. (2007) employed an automated stepwise procedure of variable selection. Testing and rejecting many variables, using stepwise regression, increase the probability of finding a significant predictor by chance, since this sifting remains undeclared (Craig *et al.*, 2007). The over-fit-

ting problem/shortfall of stepwise regression methods has also been highlighted by Babyak (2004). Furthermore, variable selection has received little attention within a geostatistical modelling framework and is usually performed as part of an explanatory analysis but carried out separately to the geostatistical model fit (Chammartin *et al.*, 2013). Also, testing and rejection of many variables increase the probability of finding a significant predictor purely by chance, while standard errors in predictive models are underestimated because this sifting process remains undeclared (Babyak, 2004; Harrell, 2001).

Taking the above into consideration, we proposed the use of the random forest (RF) algorithm, a powerful new approach developed by Leo Breiman at the University of California at Berkeley, CA, USA (Breiman, 2001) and applied for data exploration, data analysis and predictive modelling. It is a non-parametric modelling technique, which is widely used in prediction and classification problems (Garge *et al.*, 2013). Importantly, it has shown capability to deal with a large number of predictor variables and in the presence of complex interactions and highly correlated variables (Shih, 2011). In addition, RF allows predictor variables that could have been outplayed by a strong competitor to enter the ensemble. Therefore, interaction effects that would otherwise have been unnoticed are revealed (Strobl *et al.*, 2009). Finally, RF can identify relevant predictor variables by means of variable importance (VI) measures (Strobl *et al.*, 2009).

This study investigates the use of the RF algorithm to identify the most relevant and informative predictor variables from a set of candidates and provide a measure of VI within the predictive model. RF is user-friendly and the results are easy to interpret (Zhang and Bonney, 2000; Lunetta *et al*., 2004; Goldstein *et al*., 2010; Moore *et al*., 2010; Hastie *et al*., 2011). It has also been shown capable to deal with a large number of predictor variables even in the presence of complex interactions and highly correlated variables (Shih, 2011). The aim was to assess the relationship between climatic variables and malaria transmission by determining the VI of each climatic variable, as well as to evaluate what combination of climatic factors is the most associated with malaria. RF regression provides two measures of VI, the percent increase of the mean squared error (%IncMSE) and the cumulative increase in node purity (IncNodePurity) (Grömping, 2012). The %IncMSE is derived for each predictor variable from the MSE difference between the predictive measure based on the original dataset and based on a permuted dataset, where the predictor in question was randomized (Nicodemus *et al.*, 2010). This study used %IncMSE to evaluate the importance of predictor variables and ranked them in order of importance with the intention to develop predictive models of malaria transmission based on the climatic variables found to have significant measures of VI.

Materials and Methods

Malaria case data

The malaria case data used for this study were provided by the Office of Malaria Research of the South African Medical Research Council (SAMRC) from the provincial integrated malaria information system (IMIS), which is regulated by the Mpumalanga Malaria Control Program of the Department of Health. This system was developed by the SAMRC, a national research organisation in South Africa, using Microsoft Access for data entry and validation. Malaria morbidity and mortality data, consisting of both passive and active cases based on definitive diagnosis reported from December 2005 to December 2006, were provided from the IMIS for the purposes of this study. This period was chosen because it was the only one without missing data that we could be granted access to. Only malaria cases that were reported in the Ehlanzeni District were selected for analysis, as it is a malaria-prone

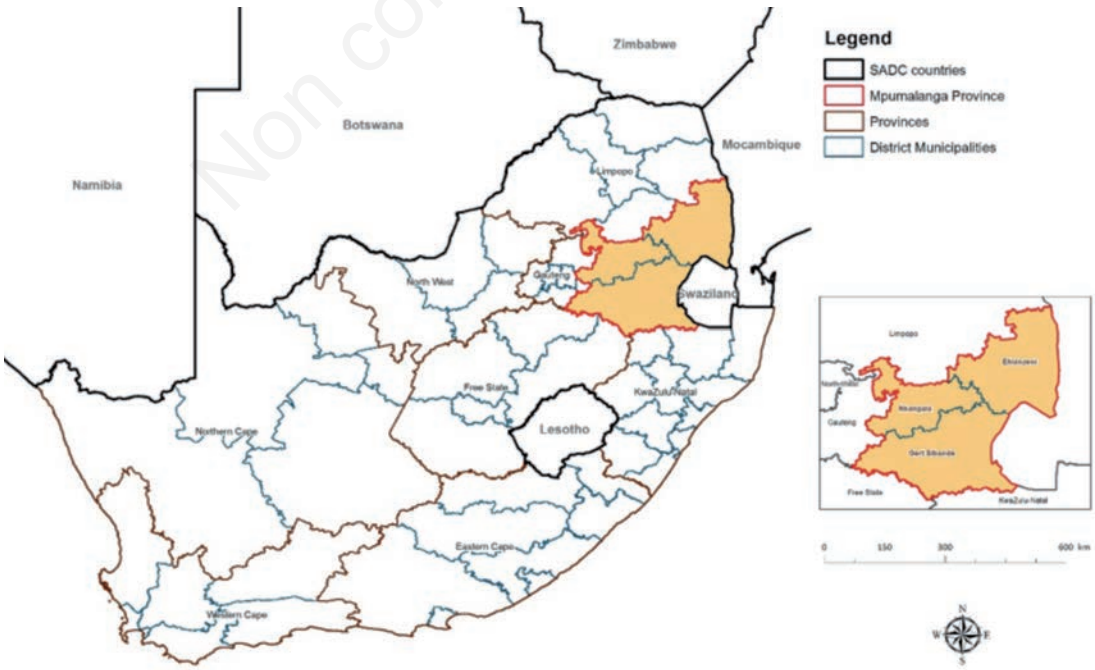

Figure 1. Map showing the location of Ehlanzeni District within Mpumalanga Province, South Africa.

District in Mpumalanga, South Africa (Figure 1). However, the district has borders both with Swaziland and Mozambique and the records therefore include many imported cases. We therefore extracted the local cases (people who could be supposed to have contracted malaria in South Africa) from the database. The data consisted of the following variables: date of diagnosis, gender, age, type of mosquito species that infected the individual, residential locality, health facility name (where the individual was screened), municipality, province and source country. Individual malaria case data were aggregated to make up a total of 396 sub-places within the study area of the Ehlanzeni District in Mpumalanga Province. South Africa is demarcated into a hierarchy of administrative boundaries, such as province, district municipality, local municipality, ward, sub-place and enumerator area. A sub-place, as defined by Statistics South Africa (Stats SA, 2011), is *the second lowest level of the place name category and the smallest administrative boundary assigned a community name and thus represents a local social boundary*.

Environmental data

The land surface temperature (LST) was obtained from the moderate resolution imaging spectroradiometer (MODIS)/Aqua LST and emissivity product (https://modis.gsfc.nasa.gov/), while the daily normalised difference vegetation index (NDVI) data were obtained from NASA's EOSDIS Reverb tool (http://reverb.echo.nasa.gov/reverb/) at a resolution of 1×1 km in hierarchal data format – earth observation system (HDF-EOS). The HDF images were then converted to GeoTIFF using the MODIS reprojection tool, v. 4.1. The raster calculator tool in ArcGIS, v. 10 (ESRI, Redlands, CA, USA) was used to apply conversion factors to LST and NDVI values as well as to convert temperature from Kelvin (K) into degrees Celsius (°C). The MODIS sensor is an instrument that records information from two satellites, one in the morning (Terra) and one in the afternoon (Aqua).

Rainfall

The monthly rainfall estimate (RFE) data was acquired from the National Oceanic and Atmospheric Administration's (NOAA) website (http://www.esrl.noaa.gov/). These data were obtained from RFE v. 2.0 implemented by the NOAA Climate Prediction Center, College Park, MD, USA. The daily data were summed up to produce 10-day totals in mm precipitation for each month at a resolution of 8×8 km in the band interleaved by line image file format (http://www.digitalpreservation.gov/formats/fdd/fdd000283.shtml). The images obtained were projected into a geographic projection and converted into GeoTIFF using ArcGIS, v.10 (ESRI).

Humidity

The monthly humidity was obtained from the NOAA website (http://www.esrl.noaa.gov/) in network common data form (netCDF) (http://www.esrl.noaa.gov/psd/data/gridded/whatsnetCDF.html) at a resolution of 0.25°×0.25°. The ArcGIS (ESRI) tool *Make NetCDF Raster Layer* was used to convert the images into GeoTIFF format.

Altitude

The altitude was obtained by creating a triangulated irregular network (TIN) in ArcGIS, v.10 (ESRI) using contour lines at 5-m intervals (http://www.planetgis.co.za). TIN is a digital data structure used in geographic information systems (GIS) for the representation of the physical land surface made up of irregularly distributed nodes with three-dimensional coordinates that are arranged in a network of non-overlapping triangles.

Lag rainfall and lag temperature

A lag time is a period between two related events; effects with a lag of 1 month for LST and rainfall were added to the climatic variables to account for possible delay of the effect of these predictor environmental variables on the number of malaria cases.

Extraction of climatic data

The ArcGIS zonal statistics tool (ESRI) was used to extract monthly averages of the climatic variables for each sub-place in the study area. The output was a comma-separated values file (http://www.computerhope.com/issues/ch001356.htm) that was imported into RStudio (https://www.rstudio.com/), which is an integrated development environment for R, a programming language and software environment for statistical computing and graphics.

Statistical methods

Random forest for regression

A regression tree is a non-parametric, statistical learning technique that is often referred to as a tree-structured algorithm (Faraway, 2005). It represents a collection of individual decision tree classifiers. The RF algorithm has become popular due to its efficiency. Moreover, it has been used in a wide variety of research studies thanks to its ability to identify important variables as well as its high accuracy of predictive models (Zhang and Bonney, 2000; Khoshgoftaar *et al*., 2007; Strobl *et al*., 2009; Goldstein *et al*., 2010; Moore *et al*., 2010). The relationship between climatic variables and malaria cases was analysed using the RF algorithm for regression based on model aggregation ideas for regression and classification problems that was developed by Breiman (2001). The RF package (http://cran.r-project.org/web/packages/randomForest/index.html) was downloaded into RStudio.

According to Breiman (2001), *each individual tree in the forest represents results from one of a set of regression trees*, each constructed based on a bootstrap sample of a dataset and a random subset of predictors. The final classification decision is the outcome of a majority vote or the weighted average of all individual trees. The importance of each predictor can also be quantified by assessing averaged prediction error across all random trees. RF is capable to deal with a large number of predictor variables even in the presence of complex interactions and highly correlated variables, and it decreases prediction errors compared to traditional regression tree methods because results are averaged among all trees (Breiman, 2001).

The RF algorithm in line with Liaw and Wiener (2002) is carried out as follows. First, draw number of trees to grow (n_{tree}) and bootstrap samples from the original data (at random with replacement) to form a training dataset. Second, for each of the bootstrap samples (training dataset), grow an *non-pruned* classification (or regression tree) with the following modification: at each node, instead of choosing the best split among all predictors, randomly sample the number of variables to select per node (m_{try}) of the predictors and choose the best split from those variables. Third, predict new data by aggregating the predictions of the majority votes for classification, average for regression, *i.e.* the n_{tree} trees. The random selection of records with replacement leaves about a third of the total number of records that are not used in the building of this tree. These records can be used to estimate the error rate of each tree, even as it is being built.

The RF algorithm not only produces predicted values but it also produces an important piece of information known as VI, which is a measure of the importance of the predictor variables. VI is a measure of the mean increase of the error of a tree, *i.e*. the MSE for regression and

misclassification rate for classification in the forest, when the observed values of this variable are randomly permuted into the so called out-of-bag (OOB) samples (Genuer *et al.*, 2010). For this study, several environmental variables – temperature, lag temperature, rainfall, lag rainfall, humidity, altitude and NDVI – were assessed in predicting malaria. The most important variables were determined by the importance scores derived from RF.

For each tree (t) the RF VI can be defined by first considering the associated OOB_t sample, then denote the error of a single tree (t) on this OOB_t sample by $errOOB_t$ followed by a random permutation of the values of X^j in OOB_t to get a perturbed sample denoted by OOB_t^j and compute $errOOB_t^j$, the error of predictor (t) on the perturbed sample:

$$VI(X^j) = \frac{1}{ntree}\sum_t \left(errOOB_t^j - errOOB_t\right)$$

When a bootstrap sample was used to construct a regression tree using environmental variables and malaria for the study period, we randomly shuffled (permuted) the values of altitude, while keeping all other variables unchanged, thus creating another regression tree using the shuffled values. Larger differences in importance scores between the models before and after permutation signified a greater importance of altitude in predicting malaria.

Averages over 50 iterations of the RF algorithm were used for each dataset because many studies (Geurts *et al.*, 2006; Genuer *et al.*, 2010; Abdulsalam *et al.*, 2011) show that this number of simulations provide stable, accurate and reliable results.

A default value for m_{try}, randomly chosen at each spilt, was used because it resulted in the lowest error estimate. The n_{tree} was set at 2000 because training runs showed it to produce a low error rate and also ensured that each and every input had to be predicted a number of times to produce reliable results. We thus used a regression RF approach based on 2000 trees and the number of variables tried at each split=1, with the call to function signified by randomForest (x=x, y=y, n_{tree}=2000, importance=TRUE).

Model validation

Unlike some machine-learning methods, RF does not require validation on a test dataset because they construct VI measures and model performances (MSE) using OOB samples, which is almost equivalent to cross-validation (Hastie *et al.*, 2001). The principle behind this cross-validation is that the dataset is split into a larger training and a smaller dataset when applied. The model is created based on the training dataset and applied to the dataset, which is also used to evaluate model performance (Walz, 2014). Therefore, when performing the RF algorithm, there is no need for cross-validation or a separate test sets to get an unbiased estimate of the test set error because it is estimated internally. However, the assessment of model fit is an important step in data analysis and the co-efficient of determination (R^2) is used to assess the quality of the fit of a linear regression model by proving an indication of the suitability of the chosen explanatory variables in predicting the response. Values for R^2 range between 0 to 1, the higher the value is, the more variability is explained by the linear regression model. Therefore model validation was performed by calculating the co-efficient of determination for each monthly predictive model generated by the RF algorithm.

Díaz-Uriarte and De Andres (2006) proposed a two-step procedure for VI selection, the first to sort variables in decreasing order of RF scores of importance and the second to cancel the variables of small importance and then perform the algorithm again for better prediction accuracy. The higher the value for %IncMSE of a predictor is, the higher the importance of that predictor in predicting the outcome (in our case, the number of malaria cases). A low importance value indicates a poor relationship between the predictor variable and the outcome because that means that permuting the attributes of that variable does not affect the predictive ability of trees on OBB samples (Bureau *et al.*, 2003).

Results

Figure 2 displays the ranking of the VI measures of the seven climatic variables in order of decreasing importance after 50 iterations of the RF algorithm.

Results averaged from 50 simulations of the RF algorithm with importance quantified as %IncMSE.

Variables with values for VI measures that are negative or close to zero are considered to be of little or no importance to the response variable (Genuer *et al.*, 2010). However, the %IncMSE values for all the climatic variables in Figure 2 were well above 0, which shows that they were significant. However, because one of the main objectives of RF variable selection is to find a small number of variables that is sufficient to allow a good prediction of the response variable, the RF algorithm was then performed on the data for each month of the year using only the 3 variables that scored the highest %IncMSE in the previous step (Figure 3).

Altitude, NDVI and temperature were the most frequently selected climatic variables for predicting malaria cases each month; these three were selected 8 times. Lag temperature and lag rainfall appeared 5 and 3 times, respectively, and humidity and rainfall both appeared twice.

Altitude was repeatedly identified as the most influential variable in predicting malaria. It was selected as the top predictor variable for four months (February, March, April and May). Lag rainfall was the top predictor for January and December, NDVI for August and September, temperature for June and November, while rainfall and lag temperature were the top predictors for only one month, July and October, respectively. Figure 4 illustrates the relationship between altitude and the occurrence of malaria cases. Altitude increased from East to West and this coincided with a decrease in the number of malaria cases in the same direction. The highest number of cases occurred at altitudes less than 600 m and decreased as the altitude increased further.

As previously stated, in addition to providing measures of VI, the RF algorithm also produces a predictive model in the form of an output of the implementation of the algorithm. Figure 5 shows the prediction performance of the top 3 statistically significant variables that were identified in the variable selection procedure. R^2, also known as the co-efficient of determination, was used to indicate the correlation between predicted malaria cases and actual malaria cases. R^2 (with a range of 0-1) is the fraction of the variability in Y that can be explained by the variability in X through their relationship.

Figures 6 and 7 illustrate the monthly spatial distribution of the observed and predicted cases of malaria in the Ehlanzeni District obtained from the predictive models. It is evident that predictions of the RF models are more accurate for sub-places with 10 or less total malaria cases. The model for October had an usually low R^2 in comparison with the other months, which is an indication of a very poor model and is likely due to the true random behaviour of the algorithm.

The RF algorithm provides an OOB error estimate, which is an internal error estimate of a random forest as it is being constructed (provided in Appendix 1). When performing the RF algorithm, after each tree of the random forest is built, the forest makes predictions on each indi-

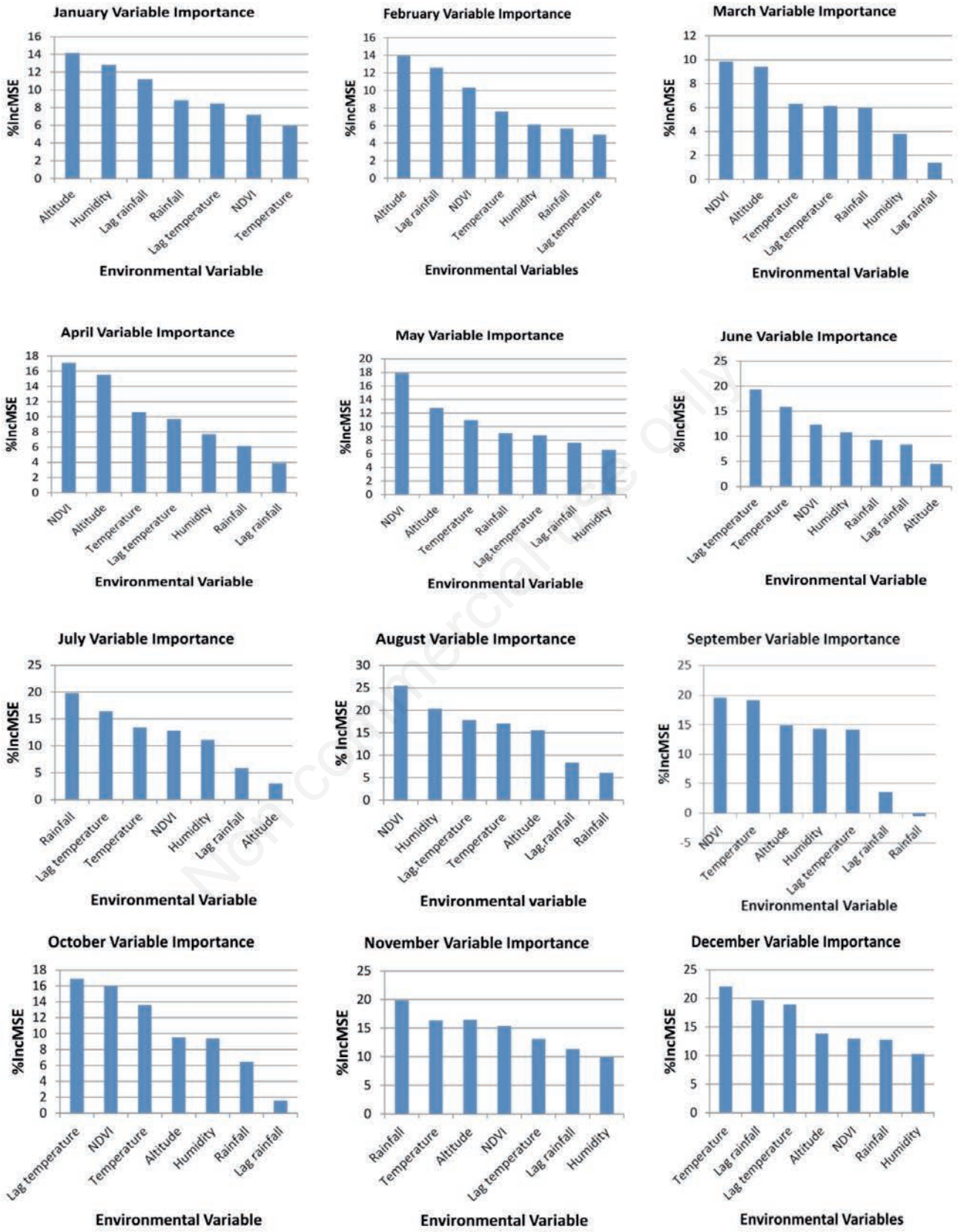

Figure 2. The variable importance measure of predictor variables for malaria.

vidual raw training dataset observation relying on the previously described technique of bagging. Once the algorithm builds the first tree in the forest it does not predict for all observations; this is because a bootstrapped observation, which was not used in the prediction, is used for those observations (Brence and Brown, 2004). The general trend observed in the graphs showing error rate is that the error rate of the predictive models starts off relatively high then drops sharply as the number of observations increases and later stabilizes as more trees are grown until the specified number of trees is grown.

Discussion

A common question that needs to be addressed in modelling is which predictor variables have the greatest influence in a predictive model and to what degree. This is the crucial issue that the RF algorithm addresses by providing individual measures of VI for several predictor variables. This study showed that altitude was the most robust predictor variable because it was selected as the single top predictor variable

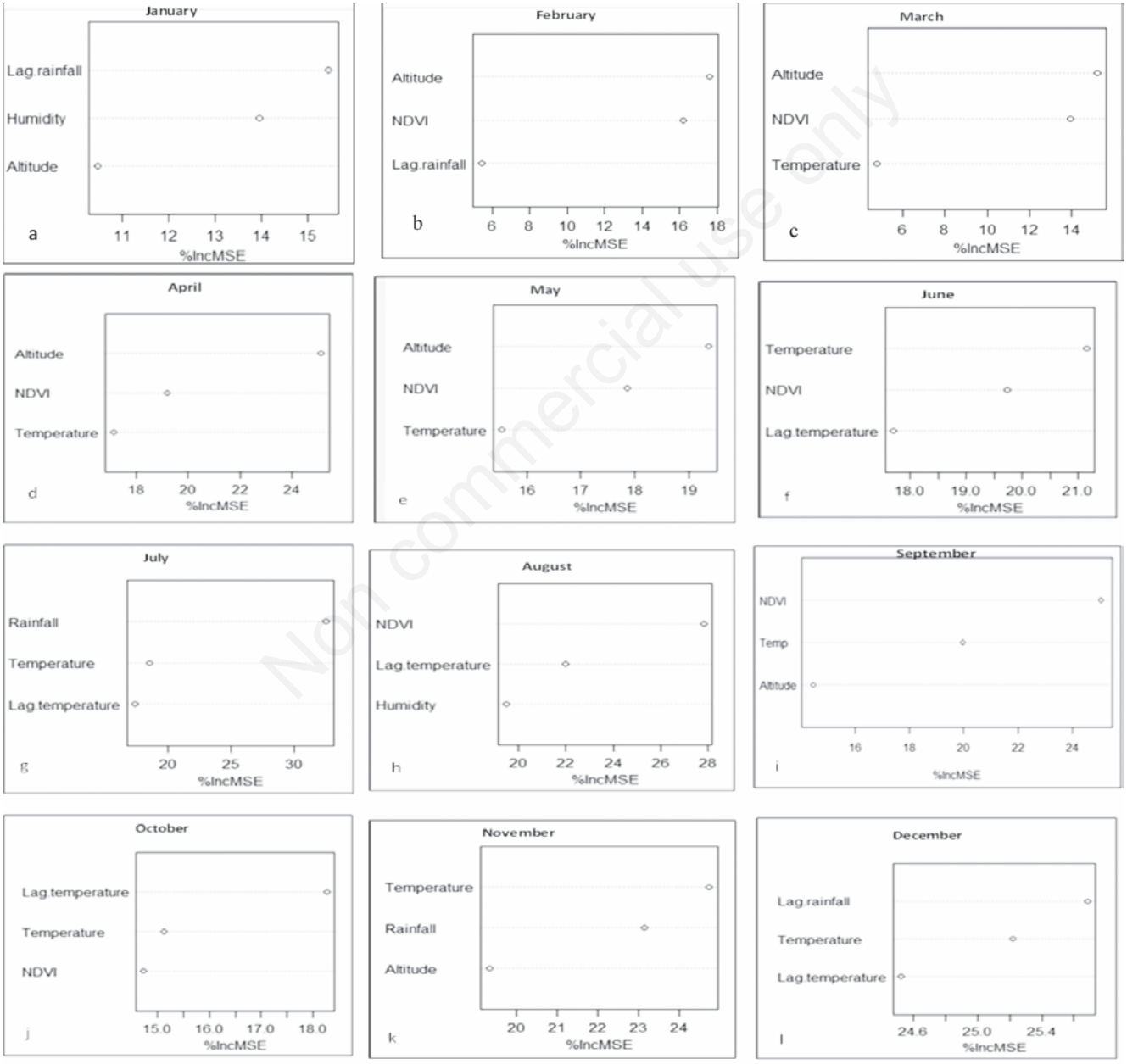

Figure 3. Variable importance measure plots generated by the random forest algorithm for the three most statistically significant variables.

for more months of the year than the other climatic variables with the %IncMSE values ranging from 16 to 26%. Ngomane *et al.* (2012) found that malaria incidence was more pronounced in the low altitude region of Ehlanzeni District in comparison to the high altitude regions of Nkangala and Gert Sibande Districts. This result suggests that altitude has an effect on malaria transmission and it echoes an investigation by Kazembe (2007) that documented a higher rate of malaria incidence at elevations below 1500 m when compared to higher altitudes. Further studies show that altitude is a key determinant of malaria risk because it restricts mosquito habitats. This is due to the findings that with every 1000 m increase in elevation, temperature decreased by 6°C. (Patz *et al.*, 2008). Craig *et al.* (2007) also found a strong positive association with malaria prevalence, they report an increase in logit(p) of 1 every 160 m. A drop in temperature results in a decline in the risk of infection because parasite development is restricted; the minimum temperature for *Plasmodium falciparum* development is said to be between 16 and 18°C (Lindsay and Martens, 1998).

Lag rainfall (16 to 26%), NDVI (26 to 28%) and temperature (21 to 26%) were the second most common predictors, each having been selected as a top predictor for 2 months. Rainfall and lag temperature were both variables that were least likely to be selected because they were the top predictors for only one month each. The results from the model validation show that the predictive models for March, April and May had the highest R^2 values (0.9023, 0.8901 and 0.8875, respectively). This indicates that the combination of climatic variables for the models for each of these 3 months (altitude, NDVI, and temperature) yielded the highest R^2 compared with other combinations and they were able to explain a high percentage of the total variation in observed malaria cases resulting in high model accuracy.

Our observations are similar to research conducted by Craig *et al.* (2007), who developed a malaria risk map using a systematic variable selection process. They found that out of 50 potential explanatory variables, rainfall, temperature and altitude were the most plausible predictors of historical malaria risk in Botswana. The only difference is that our study found NDVI to be significant in place of rainfall. Our results are also in accordance with a study by Sinka *et al.* (2010) who used RF for classification to map the prediction of dominant vectors of human malaria. They found that in the ranking of various climatic variables, altitude and temperature were the most influential in predicting malaria due to their high %IncMSE values.

Rainfall had the highest measure of importance because it had the highest %IncMSE (about 34%); therefore it was the climatic variable that made the greatest contribution to the prediction of malaria in the monthly models. This is an important finding because rainfall and temperature have proved to be two of the major environmental variables triggering malaria epidemics in warm semi-arid and high altitude areas because epidemics occur in these regions after excessive rain or increases in temperature (Ceccato *et al.*, 2012). Rainfall was followed by NDVI and lag rainfall with %IncMSE values of 28 and 26%, respectively.

Multiple iterations of the RF algorithm (50 for each month) were implemented because the reliability and stability of VI measures are greatly increased when multiple simulations are executed (Goldstein *et al.*, 2010). Apart from determining VI measures, the RF algorithm produced models that allowed for the production of predictive maps. The second aim of this study was to produce predictive maps from the pre-

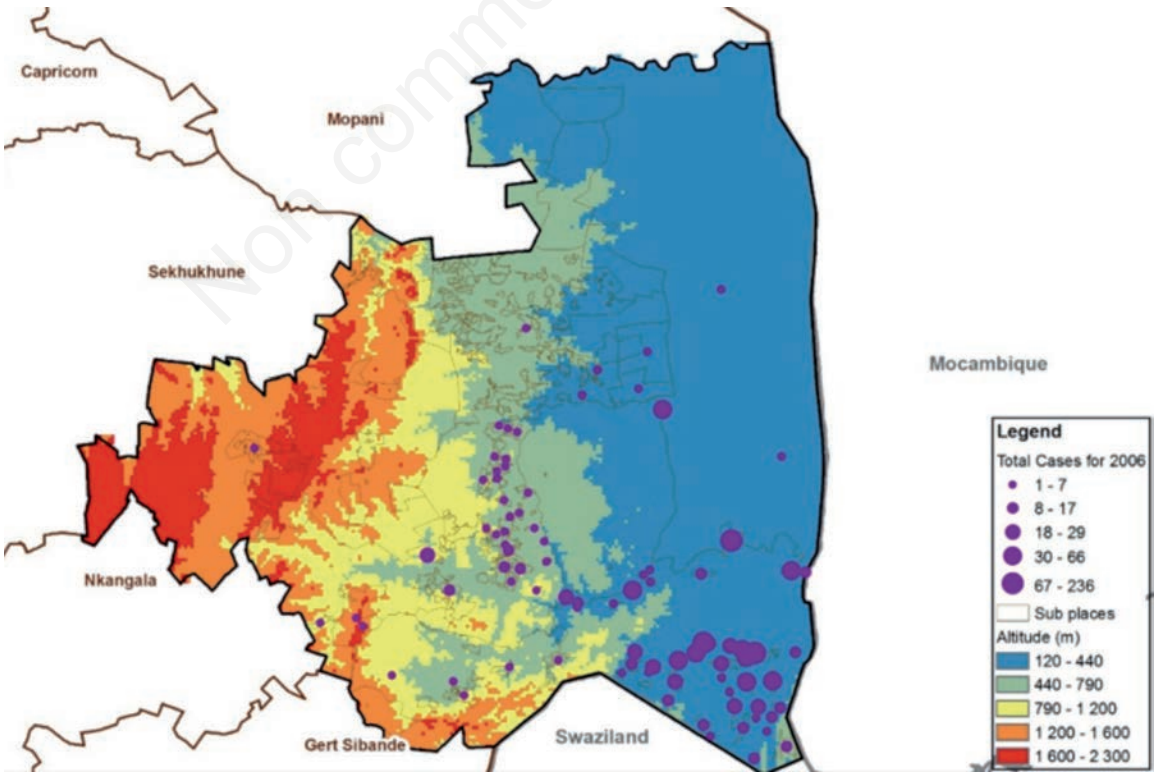

Figure 4. Association between altitude and number of malaria cases.

dicted cases output that could be used to monitor the progression of malaria and assist in targeted intervention and prevention efforts. Maps showing the distribution of the predicted number of malaria cases against the actual observed numbers also serve as a tool for disease surveillance because areas at risk of potential outbreaks can be easily identified (Hay *et al.*, 2013).

We also showed that the models generated by the RF algorithm predicted low numbers of malaria more accurately than higher numbers. Appendix 2 shows that the majority of the observed numbers of total malaria cases per sub-place were 10 and below, and the predictive maps show that the models were able to predict fairly accurately within the range of 0-10 observed cases as seen on the maps for the months of June, July, August, September and December. However, the RF algorithm does not predict extreme values accurately (Horning, 2013) and it usually considers values greater than 10 as outliers (Breiman, 2002). Therefore, as shown in the maps, the predictive capabilities of the

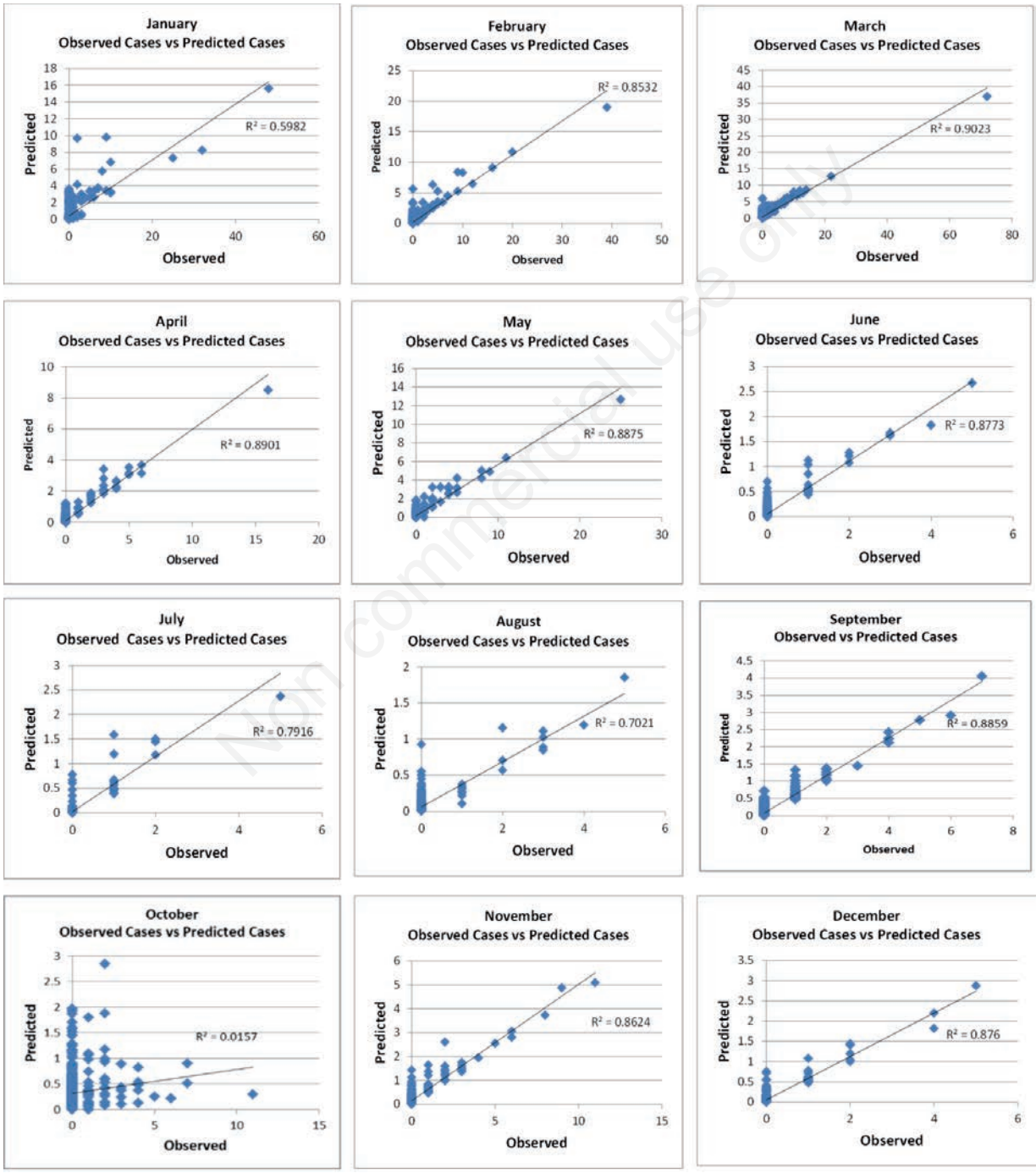

Figure 5. The coefficient of determination (R^2) of observed and predicted cases.

models are less accurate for high-observed numbers as seen for the months of January, February and March and May.

Correlation was not addressed separately outside the RF algorithm in this study because instead of searching over all the variables at each node for the optimal split, the RF algorithm reduces correlation among trees by performing a search over a random subset of data at each node. It continues to split the data until no further splits possible, RF leaves the tree *non-pruned* because bagging is used to decrease variance resulting from the lack of pruning (Goldstein *et al.*, 2010). Also, a study by Genuer *et al.* (2010) showed that even in the presence of highly correlated variables, the fine tuning parameter m_{try} and selecting of a sufficient number of trees ensured that VI measures produced with different iterations of the algorithm do not vary systematically and RF is able to select significant variables by allocating them high measures of importance and allocating variables of low importance with low measures of importance.

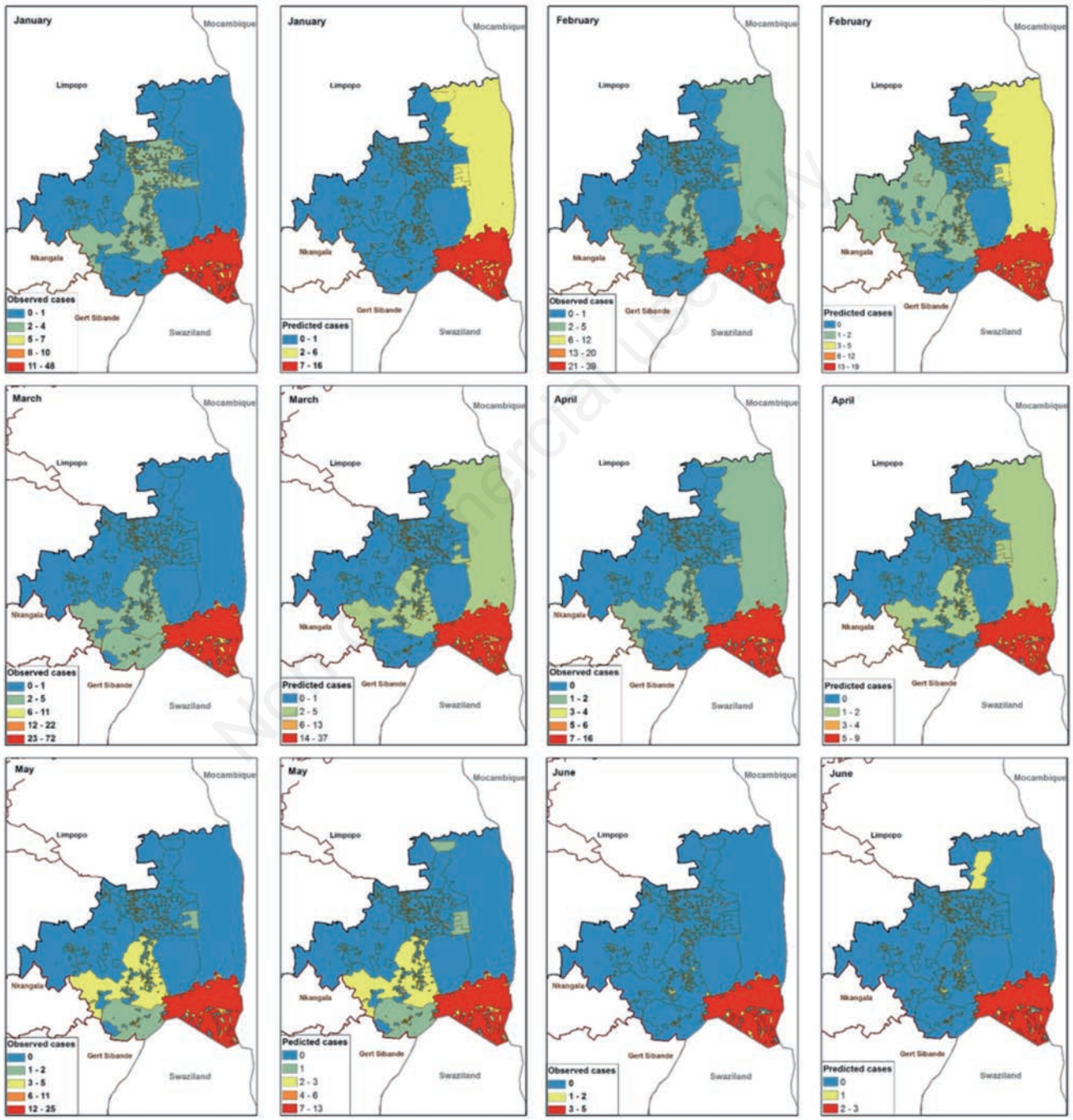

Figure 6. Maps showing observed and predicted cases of malaria by sub-place as obtained by the application of the random forest models (January to June).

One of the limitations we encountered is that the RF algorithm cannot predict beyond the range of data in the training data; thus, the more data is available, the greater and more reliable the prediction ability of predictive models will become. In light of this, a follow up to this study would be able to obtain malaria case data and climatic data over a longer period of time, such as the past decade, and obtain measures of VI for all the climatic variables for each month over each of the years to evaluate the pattern of statistically significant climatic variables most associated with cases of malaria. Also, there is a presence of a high number of zeroes in the data; however, these zeroes are true because they represent the absence of malaria cases in those sub-places. For this reason, they cannot be excluded from the RF algorithm. Cameron and Trivedi (2013) state that real-life data frequently display over dispersion through excess zeroes and a zero value has special appeal in many situations because it divides the population into subpopulations in a meaningful way. If we apply this statement to this study, zeroes in

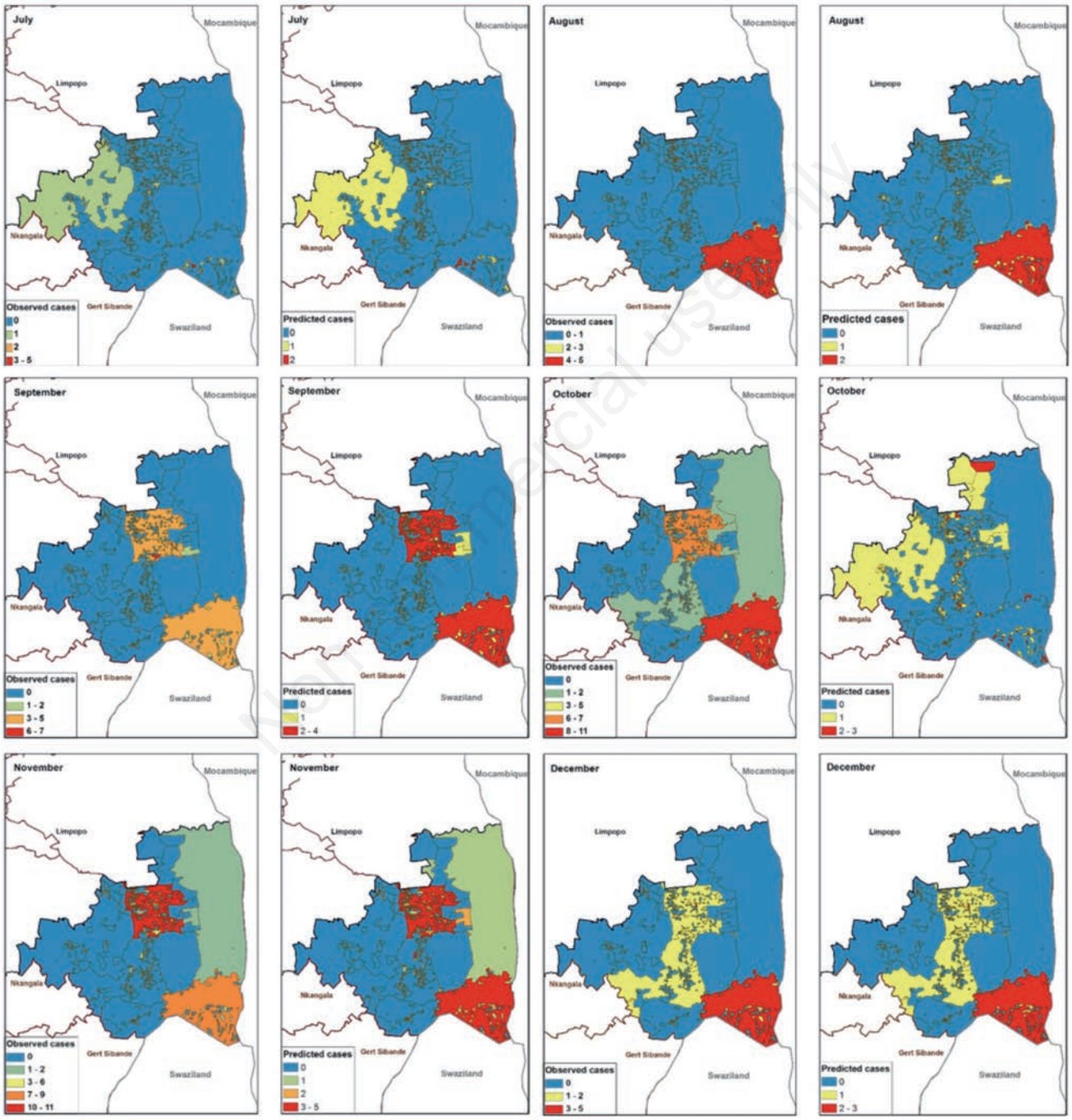

Figure 7. Maps showing observed and predicted cases of malaria by sub-place as obtained by the application of the random forest models (July to December).

the dataset of malaria cases are important because they separate sub places that have actual observed malaria cases and those that do not have any cases.

Conclusions

The RF algorithm can be used as a tool to filter out irrelevant predictor variables and focus on those that have the greatest impact on malaria. This can be viewed as a time saving exercise for data collection and experimental run time because for future studies, efforts can be concentrated towards only collecting data for significant climatic variables.

Altitude was the most robust climatic predictor influencing malaria and the most frequently selected variable followed by lag rainfall, NDVI and temperature (each selected as the dominant variable for 2 months). Rainfall and lag temperature were each selected as dominant predictor variables, but only for one month each.

Predictive models for the months of March, April and May had the highest model accuracy due to the fact that the combination of climatic factors produced the highest R^2 values. All three months had the same top three statistically significant climatic variables that were ranked in the same order of importance: altitude was first followed by NDVI and temperature was last.

There is potential for the predictive maps generated from the prediction capability of the RF algorithm to be used in improving public health as an operational malaria EWS by determining areas that are at risk of high numbers of malaria cases in the future as well as by aiding the efficient targeting of intervention and prevention measures through focusing on areas that are at risk of outbreaks.

References

Abdulsalam H, Skillicorn DB, Martin P, 2011. Classification using streaming random forests. IEEE T Knowl Data En 23:22-36.

Babyak MA, 2004. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med 66:411-21.

Breiman L, 2001. Random forests. Mach Learn 45:5-32.

Breiman L, 2002. Manual on setting up, using, and understanding random forests v3. 1. Statistics Department, University of California, Berkeley, CA, USA.

Brence JR, Brown DE, 2004. Analysis of robust measures for random forest regression. University of Virginia, Charlottesville, VA, USA.

Brooker S, Clarke S, Njagi JK, Polack S, Mugo B, Estambale B, Muchiri E, Magnussen P, Cox J, 2004. Spatial clustering of malaria and associated risk factors during an epidemic in a highland area of western Kenya. Trop Med Int Health 9:757-66.

Bureau A, Dupuis J, Hayward B, Falls K, Van Eerdewegh P, 2003. Mapping complex traits using Random Forests. BMC Genet 4(Suppl.1):S64.

Cameron AC, Trivedi PK, 2013. Regression analysis of count data. Cambridge University Press, Cambridge, UK.

Castillo-Riquelme M, Mcintyre D, Barnes K, 2008. Household burden of malaria in South Africa and Mozambique: is there a catastrophic impact? Trop Med Int Health 13:108-22.

Ceccato P, Vancutsem C, Klaver R, Rowland J, Connor SJ, 2012. A vectorial capacity product to monitor changing malaria transmission potential in epidemic regions of Africa. J Trop Med 2012:595948.

Chammartin F, Hürlimann E, Raso G, N'goran EK, Utzinger J, Vounatsou P, 2013. Statistical methodological issues in mapping historical schistosomiasis survey data. Acta Tropica 128:345-52.

Craig MH, Sharp BL, Mabaso ML, Kleinschmidt I, 2007. Developing a spatial-statistical model and map of historical malaria prevalence in Botswana using a staged variable selection procedure. Int J Health Geogr 6:44.

Craig MH, Snow RW, Le Sueur D, 1999. A climate-based distribution model of malaria transmission in Sub-Saharan Africa. Parasitol Today 15:105-11.

Díaz-Uriarte R, De Andres SA, 2006. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 7:1.

Faraway JJ, 2005. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. CRC press, Boca Raton, FL, USA.

Garge NR, Bobashev G, Eggleston B, 2013. Random forest methodology for model-based recursive partitioning: the mobForest package for R. BMC Bioinformatics 14:125.

Genuer R, Poggi J-M, Tuleau-Malo TC, 2010. Variable selection using random forests. Pattern Recogn Lett 31:2225-36.

Gerritsen A, Kruger P, Van Der Loeff M, Grobusch MP, 2008. Malaria incidence in Limpopo Province, South Africa, 1998-2007. Malaria J 7:162.

Geurts P, Ernst D, Wehenkel L, 2006. Extremely randomized trees. Mach Learn 63:3-42.

Goldstein BA, Hubbard AE, Cutler A, Barcellos LF, 2010. An application of Random Forests to a genome-wide association dataset: methodological considerations & new findings. BMC Genet 11:49.

Gosoniu L, Vounatsou P, Sogoba N, Smith T, 2006. Bayesian modelling of geostatistical malaria risk data. Geospat Health 1:127-39.

Govere J, Durrheim D, Coetzee M, Hunt R, La Grange J, 2000. Captures of mosquitoes of the Anopheles gambiae complex (Diptera: Culicidae) in the Lowveld region of Mpumalanga Province, South Africa. Afr Entomol 8:91-9.

Grömping U, 2012. Variable importance assessment in regression: linear regression versus random forest. Am Stat 63:308-19.

Harrell FE, 2001. Regression modeling strategies: with applications to linear models, logistic regression, and survival analysis. Springer, Amsterdam, The Netherlands.

Hastie T, Tibshirani R, Friedman JH, 2001. The elements of statistical learning: data mining, inference, and prediction. Springer, New York, NY, USA.

Hastie TJ, Tibshirani RJ, Friedman JH, 2011. The elements of statistical learning: data mining, inference, and prediction. Springer, Amsterdam, The Netherlands.

Hay SI, Battle KE, Pigott DM, Smith DL, Moyes CL, Bhatt S, Brownstein JS, Collier N, Myers MF, George DB, Gething PW, 2013. Global mapping of infectious disease. Philos T Roy Soc B 368:20120250.

Horning N, 2013. Introduction to decision trees and random forests. American Museum of Natural History's, New York, NY, USA.

Karthe D, 2010. Geographic determinants of malaria transmission. A Case Study from Kossi Province, Burkina Faso. Available from: https://ediss.uni-goettingen.de/bitstream/handle/11858/00-1735-0000-000D-F177-A/karthe.pdf?sequence=1

Kazembe LN, 2007. Spatial modelling and risk factors of malaria incidence in northern Malawi. Acta Tropica 102:126-37.

Khoshgoftaar TM, Golawala M, Hulse JV, 2007. An empirical study of learning from imbalanced data using random forest. In: IEEE Computer Society Washington, ed. Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence, 2007 October 29-31, Patras, Greece. IEEE Computer Society Washington,

DC, USA, pp 310-7.
Kleinschmidt I, Bagayoko M, Clarke G, Craig M, Le Sueur D, 2000. A spatial statistical approach to malaria mapping. Int J Epidemiol 29:355-61.
Liaw A, Wiener M, 2002. Classification and regression by randomForest. R News 2:18-22.
Lindsay S, Martens W, 1998. Malaria in the African highlands: past, present and future. B World Health Organ 76:33.
Lunetta KL, Hayward L, Segal J, Van Eerdewegh P, 2004. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 5:32.
Mabaso ML, Craig M, Ross A, Smith T, 2007. Environmental predictors of the seasonality of malaria transmission in Africa: the challenge. Am J Trop Med Hyg 76:33-8.
Moore JH, Asselbergs FW, Williams SM, 2010. Bioinformatics challenges for genome-wide association studies. Bioinformatics 26:445-55.
Ngomane L, De Jager C, 2012. Changes in malaria morbidity and mortality in Mpumalanga Province, South Africa (2001-2009): a retrospective study. Malaria J 11:19.
Nicodemus KK, Malley JD, Strobl C, Ziegler A, 2010. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics 11:110.
Ostfeld RS, Glass GE, Keesing F, 2005. Spatial epidemiology: an emerging (or re-emerging) discipline. Trends Ecol Evolut 20:328-36.
Patz JA, Olson SH, Uejio CK, Gibbs HK, 2008. Disease emergence from global climate and land use change. Med Clini North Am 92:1473-91.
Shih S, 2011. Random forests for classification trees and categorical dependent variables: an informal quick start R guide. Stanford University, Stanford, CA, USA.
Silal SP, Barnes KI, Kok G, Mabuza A, Little F, 2013. Exploring the Seasonality of reported treated malaria cases in Mpumalanga, South Africa. PLoS ONE 8:e76640.
Sinka ME, Rubio-Palis Y, Manguin S, Patil AP, Temperley WH, Gething PW, Van Boeckel T, Kabaria CW, Harbach RE, Hay SI, 2010. The dominant Anopheles vectors of human malaria in the Americas: occurrence data, distribution maps and bionomic précis. Parasites Vectors 3:72.
Stats SA, 2011. Census 2011 statistical release. Statistics South Africa, Pretoria, South Africa.
Strobl C, Malley J, Tutz G, 2009. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods 14:323.
Walz Y, 2014. Remote sensing for disease risk profiling: a spatial analysis of schistosomiasis in West Africa. Julius-Maximilians-Universität Würzburg, Würzburg, Germany.
Zhang H, Bonney G, 2000. Use of classification trees for association studies. Genet Epidemiol 19:323-32.