Published in final edited form as:
Acta Trop. 2011 July ; 119(1): 30–37. doi:10.1016/j.actatropica.2011.04.003.

Analysing the generality of spatially predictive mosquito habitat models

Li Li[a,*], **Ling Bian**[b], **Laith Yakob**[c], **Guofa Zhou**[d], and **Guiyun Yan**[d]

[a]School of Environmental Science, Murdoch University, Murdoch, Western Australia, Australia

[b]Department of Geography, University at Buffalo, Amherst, New York, USA

[c]School of Biological Sciences, xUniversity of Queensland, Brisbane 4072, Australia

[d]Program in Public Health, College of Health Sciences, University of California, Irvine, CA, USA

Abstract

The increasing spread of multi-drug resistant malaria in African highlands has highlighted the importance of malaria suppression through vector control. Its historical success has meant that larval control has been proposed as part of an integrated malaria vector control program. Due to high operation costs, larval control activities would benefit greatly if the locations of mosquito habitats could be identified quickly and easily, allowing for focal habitat source suppression. Several mosquito habitat models have been developed to predict the location of mosquito habitats. However, to what extent these models can be generalised across time and space to predict the distribution of dynamic mosquito habitats remains largely unexplored. This study used mosquito habitat data collected in six different time periods and four different modelling approaches to establish 24 mosquito habitat models. We systematically tested the generality of these 24 mosquito habitat models. We found that although habitat–environment relationships change temporally, a modest level of performance was attained when validating the models using data collected from different time periods. We also describe flexible approaches to the predictive modelling of mosquito habitats, that provide novel modelling architecture for future research efforts.

Keywords

Model generality; Spatial predictive habitat models; Temporal generality; Spatial generality; Malaria; Mosquito; Larval habitat

1. Introduction

Malaria control efforts have increased considerably in African highland areas (Malakooti et al., 1998; Shanks et al., 2000; Akhwale et al., 2004) in order to manage recent fatal malaria outbreaks. However, commonly used malaria control tools, such as anti-malaria drugs, have become problematic, due to the spread of multi-drug resistant malaria. Consequently, the importance of transmission reduction through vector control has been highlighted (http://www.rti.org, Killeen et al., 2002a). Malaria vector control has demonstrated historical success in several countries (Utzinger et al., 2001; Killeen et al., 2002b). Some theoretical

*Corresponding author. Tel.: +86 10 58748766; fax: +86 10 58748230. lili36@ub-alumni.org (L. Li).

Conflicts of interests
The authors declare that they have no conflicts of interests.

studies have demonstrated that vector control via habitat source suppression can be an effective supplement to the use of Insecticide-Treated-Nets (Gu and Novak, 2005; Yakob and Yan, 2009, 2010).

A major drawback of vector control strategies, however, is their costs. Vector control activities would benefit greatly if the locations of mosquito habitats could be identified quickly and easily, enabling a more focal habitat source suppression. Species habitat models have been increasingly used in species management and conservation, informing the targeting of species management locations in order to reduce negative environmental impacts. Intuitively, the success of focal habitat management is dependent on the predictive power of habitat models. Several mosquito habitat models have been developed to predict the location of mosquito habitats (Hay et al., 1998; Brownstein et al., 2002). However, because species distribution models are often highly dependent on field observations (Guisan and Zimmermann, 2000), the development of such models is often costly, requiring both extensive field work and experts with data analysis abilities.

Empirical models based on data from a particular time may lose much of their predictive power when applied to data at different time points due to the following two reasons. First, the approach used for the model development may not be able to accommodate the temporal changes in species–environment relationships (Bulluck et al., 2006). Second, datasets collected at a particular time may not be representative of the full spectrum of habitat conditions (Strauss and Biedermann, 2007).

The generality of habitat models requires thorough assessment, prior to making general inferences or predictions (Strauss and Biedermann, 2007). Whether mosquito habitat models based on data collected in a particular time can be used to predict habitat occurrence at a different time is poorly understood.

In this study, we first introduce several commonly used habitat modelling approaches. We then test the generality of these habitat models using mosquito habitat data collected in different time periods. Finally, we discuss the impact of generality of mosquito habitat models in strategy development for malaria control.

2. Background

The goal of most habitat models is to differentiate land-types based on their suitability to a particular species. This goal is typically achieved by using statistical tools to either compare the environmental conditions in habitat-present area with such conditions in habitat-absent areas or with conditions found throughout the entire study area. The former approach is referred to as a presence–absence model and it requires data on both the presence and absence of habitats. The latter requires only data on the presence of habitats and is referred to as a profile-type model (Hirzel et al., 2002). We now describe the mathematical details of both of these approaches.

2.1. Presence–absence models

The most popular presence–absence model, logistic regression (LR) analysis, is a generalised linear model and is designed to analyse binary data. It has been successfully applied to model the habitats of various species and is the most commonly used approach for habitat modelling (Morrison et al., 1998). A generalised LR function is shown in Eq. (1).

$$\ln\left(\frac{p}{1-p}\right)=\beta_0+\beta_1\times x_1+\cdots+\beta_n\times x_n \quad (1)$$

where p is the probability of the occurrence of habitats, β_0 is a constant, and $\beta_1, \ldots, \beta_n$ are coefficients associated with the environmental variables $x_1 \ldots, x_n$.

When LR is applied to spatial data, however, the problem of spatial autocorrelation often occurs (Guisan and Zimmermann, 2000). Similar to most of the regression techniques, it is assumed that the individual observations are independent from each other. However, spatial autocorrelation frequently occurs for many ecological phenomena, possibly resulting in unreliable parameter estimation or inefficient estimates for LR. A simple approach to dealing with spatial autocorrelation in observations is to selectively remove some samples to reduce the level of spatial autocorrelation in the data (Bian et al., 2006). However, certain aspects of observation, for example, seasonality, can be removed during the thinning process. Spatial regression approaches, such as spatial autoregressive modelling, can take into account this dependence in habitat observations without the thinning process (Mertens and Lambin, 1997). Spatial autoregressive models that deal with a binary dependent variable are often referred to as spatial logistic regression (SLR) models. The generic equation for a SLR is given in Eq. (2):

$$y = \rho \times W \times y + \beta_0 + \beta_1 \times x_1 + \cdots + \beta_n \times x_n \quad (2)$$

where y is equal to $\ln(p/(1-p))$, W is the neighbourhood relationship continuity matrix and ρ is a parameter that reflects the strength of spatial dependency between the elements of the dependent variable. The term, $\rho \times W \times y$, is incorporated to correct the error introduced by the spatial autocorrelation in the dependent variables. As a consequence of introducing this term, the residual errors become independent.

In both LR and SLR, environmental variables are linearly linked to the dependent variable via a link function (e.g. the right side of Eqs. (1) and (2)). This modelling structure is not particularly flexible and cannot depict the non-linear relationships between dependent and independent variables.

A modelling approach that is considered more flexible than LR is artificial neural network (ANN). ANNs have been used extensively in artificial intelligence (AI) (Rumelhart et al., 1986). Using non-linear statistical approaches, ANNs adapt the structure of connectivity of a modelled system such as neuronal networks. They learn by adjusting the weights on its neurons to minimise error. This approach is considered a promising area of predictive habitat distribution modelling, since ANN models can be non-linear and species–environment relationships can often be non-linear (Guisan and Zimmermann, 2000; Recknagel, 2001). Some studies have shown that, for non-linear relationships and interactions among variables, neural networks may be better predictors than the generalised linear statistical models. These studies were mainly based on the principle of a feed-forward ANN with back-propagation (Rumelhart et al., 1986). The back-propagation approach is the most widely used approach for ANN training (Lek and Guagan, 1999).

A back-propagation network typically consists of three neuron layers: an input layer, one or more hidden layers and an output layer each including one or more neurons, as shown in Fig. 1. Each input neuron represents an environmental variable. Environmental data observed at habitat-present locations and habitat-absence locations enter the hidden layers. In the hidden layers, environmental data are summed up and then fed into an activation function, which generates the output of the patch. The connections between the input patch and the hidden layers are expressed as an interconnection weight. Through a learning algorithm, the weights are adjusted iteratively, increasing the agreement between the observed and predicted presence of the species (Lek and Guagan, 1999). The output layer

only has one criterion, indicating the presence or absence of a mosquito habitat. Back-propagation is a non-parametric approach, making very few assumptions about the data.

A major limitation for presence–absence models is that the data on the absence of habitats are often difficult to obtain (Hirzel et al., 2001). The common solution to this problem is to simply generate pseudo-absence data by selecting locations randomly from the areas where habitats are not observed. The pseudo-absence data are then used to represent the environmental conditions that are unfavourable to habitats. In the following section, the process by which pseudo-absence data was generated as part of the construction process of our models will be detailed.

2.2. Profile-type models

Reliance of habitat modelling approaches on both presence and absence data could be troublesome for those studies which have difficulties in obtaining absence data (Hirzel et al., 2002). Profile-type models were developed to deal with this problem. These models assume that habitats are non-randomly distributed regarding environmental variables. For example, mosquito habitats are expected to occur preferentially in locations with optimal water availability. The optimal conditions for habitats may be quantified by comparing the water availability of locations in which the species was observed with that of the whole study area. The water availability at these two types of locations may differ with respect to its mean and its variances.

The most popular profile-type model is environmental niche factor analysis (ENFA). It identifies optimal habitat conditions by computing habitat suitability indexes using environmental variables and presence-only data (Hirzel et al., 2002). It first transforms environmental variables into a set of uncorrelated factors by using an approach similar to principle component analysis (PCA). The first factor is chosen to maximise the variation between a “global” space and a “focal” space. The global space represents the collective environmental condition of an entire area, and is defined by the original variables. The focal space represents the most preferred habitat condition within the area, and is a part of the global space. The first factor thus represents the species marginality. Each sequential factor maximises the ratio between the variance of the global space and that of the focal space remaining in the data (Hirzel et al., 2002). These factors represent the specialization of the niche. After obtaining these factors, ENFA develops a suitability index for each location by calculating a combination of the values of each factor on a per-location basis. In order to account for the differential ecological importance of the factors, equal weighting was given to marginality and specialization. Since the entire marginality component falls within the first factor, the specialization component is apportioned among all factors proportionally, according to their eigenvalue (Hirzel et al., 2002).

Although all four of the aforementioned approaches have achieved considerable success in predicting habitats for different species, many questions remain regarding the generality of habitat models. It is unknown whether and how presence-only or the pseudo-absence data affect the generality of habitat models. The extent to which spatial autocorrelation in data and the flexibility of habitat approaches affect the generality of models is also unknown.

3. Methods

3.1. Study area, training data and validation data

The study site is a 4 × 4 km area in Kakamega district, western Kenya. The hilly landscape of the study area is typical of East African highlands. The African highlands are characterized by alternating rainy and dry seasons with a clear inter-annual variation in precipitation (Minakawa et al., 2001; Boken et al., 2005). The rainy season usually starts in

April and lasts 2–3 months. The dry season starts in December and ends in March (Minakawa et al., 2002). February and May are the most representative months of the dry and rainy seasons, respectively.

Data availability is a major constraint of mosquito habitat modelling in African countries. This study used only one data source on the environment, the most recent aerial photography survey maps. A 20 m interval contour map was digitised from the survey map with a 1:50,000 scale. This contour map was digitised to derive a digital elevation model (DEM) with 30 m resolution. It is common practice to derive a DEM with a 30 m resolution from maps with a scale of 1:50,000 (Defense Mapping Agency, 1980). Because mosquito larvae are constrained by aquatic environments, a stream map was also digitised from the survey maps and verified by field surveys (Minakawa et al., 2001). Five environmental variables were then derived from this DEM to describe the aquatic environments in the study area. Three variables were used to represent surface water availability: curvature, distance to streams and wetness index. Wetness index indicates soil moisture level in relation to water flow patterns, which is calculated based on the local upslope contributing area and slope. Two variables were used to represent habitat conditions: elevation and heatload index. Heat-load index presents micro-scale solar radiation variation, which is calculated based on latitude, slope, and aspect. The definitions of these five variables and sources are provided in Table 1.

Anopheles gambiae sensu stricto is the primary malaria vector species in western Kenya highlands, constituting more than 95% of malaria-harbouring mosquitoes (Minakawa et al., 2002). Thorough searches of all identifiable *An. gambiae* habitats (referred to from here on as mosquito habitats) were conducted in weeks 1 and 2 of February and May in 2003, 2004 and 2005, generating a total of six point maps of aquatic habitat locations. Note that although mosquito habitats refer to discrete water bodies in mosquito ecology, this study adopts a ecological habitat modelling framework and mosquito habitats here refer to locations where mosquito larvae are found. Detailed information on the habitat identification approach can be found in Li et al. (2009). Only the data on the presence of *An. gambiae* larvae were used in this study, since the quantification of larval abundance is prone to sampling errors, particularly in large aquatic environments (Chubachi, 1976). The numbers of habitats observed in these six field trips were 301, 721, 201, 416, 77 and 410 in chronological order. Based on the field observations, 99.9% of habitats had a diameter of <50 m. For habitat modelling purposes, 721 locations that were at least 100 m away from all habitat-presence locations were randomly selected in the study area and these 721 locations were used to present habitat pseudo-absence locations. To prepare data for the analysis, we first split each of the six habitat datasets into two parts: 2/3 for model “training” (hereafter referred to as training data) and 1/3 for model validation or “testing” which are not used in model construction (hereafter referred to as validation data). To prepare data for LR, SLR and ANN, each of the training parts were integrated with an equal numbers of locations selected from the habitat pseudo-absence locations. For example, for data collected in February 2003 (301 habitat locations, shown in Fig. 2), we first randomly selected 2/3 of habitat-presence locations, which resulted in 200 locations. We then randomly selected 200 locations from habitat-absence locations (721 locations). We integrated these 200 habitat-presence locations with 200 habitat-absence locations and created a training dataset consisting of 400 samples.

To implement SLR and ANN approaches, there was a need to specify several additional parameters. To use the SLR approach, it is necessary to calculate a weight matrix (the W term in Eq. (2)) for the dependent variable. The first-order and second-order contiguity spatial weight matrices are a recommended method for spatial regression analysis (LeSage and Pace, 2004). In the first-order contiguity matrix, two geographic location i and j are

neighbors if directly share a border. The second-order contiguity matrix considers two geographic units *i* and *j* are neighbors if they directly share a border or if they have a common neighbor with which they directly share a border. To calculate contiguity matrix, point data are usually converted into teisson polygons. These two matrices were both created for habitat data and, based on a preliminary analysis, the models using the first-order contiguity spatial weight matrix were found to have a better prediction accuracy. In this study, the first-order contiguity spatial weight matrix was used. For the SLR approach, we selected a Bayesian model fitting approach using a truncated normal prior. It is the recommended approach for model fitting as it takes into account the spatial uncertainty of the data (LeSage and Pace, 2009). For the ANN model fitting, we selected a probit link function. This is the link function recommended for binary output data (Huettmann and Linke, 2003).

Using four approaches and six training datasets, we generated 24 models. ArcMap was used for data manipulation (http://www.esri.com). ArcGIS extensions that are used to derive the environmental variables are listed in Table 1. ENFA was implemented using Biomapper 4.0 (http://www2.unil.ch/biomapper). LR, SLR and ANN were carried out in a MATLAB environment. The MATLAB extensions used in this study include the *Neural Network Toolbox 6.0* (http://www.mathworks.com/products/neuralnet/) and the *Econometrics Toolbox* (http://www.spatial-econometrics.com).

3.2. Assessing the accuracy and the temporal generality of models

There are several metrics that can be used to evaluate model performance, for example, *R* square. These metrics typically examine how well a model was fitted to its training dataset. In the current study, the major concern is the generality of habitat models, hence cross-validation was used. In cross-validation for habitat models, models are first applied to data collected at both habitat-presence and habitat-absence locations. The percentages of correctly predicted habitat-presence and habitat-absence locations are calculated. The models with the highest accuracies are considered the model with the strongest generality. In this study, the habitat-absence locations are not based on the field observations and they are selected based on the assumption that the absence locations are 100 m away from the habitat-presence locations. To ensure the robustness of the model validation, the datasets used in the validation are from the field observations and pseudo-absence locations were not used. To test whether the 24 predictive maps fitted the training data from which they were calculated, each of them was compared with their training dataset (excluding the habitat-absence locations). To test whether the 24 models could be applied to the locations that were not utilised in generating the training dataset, each of them was applied to its testing set (hereafter referred to as testing data in the same time period). The prediction accuracy based on training datasets could be biased, since models could be overly fitted to the data. In comparison, the prediction accuracy based on testing data can be more reliable in defining the predictive power of models. To examine whether these 24 maps could be applied to other time periods, each of them was applied to five other habitat datasets collected in different time periods (hereafter referred to as testing data in different time periods). Therefore, each of the 24 maps was applied to seven datasets. The outputs are values ranging from 0 to 1, denoting the probability of occurrence. To decide whether habitats occur in a location, a cut-off point for the probability of occurrence has to be chosen. If the probability of occurrence of a location is higher than this cut-off point, this location is considered a predicted habitat location. In this study, if a predicted value was greater than 0.5, it was considered a predicted habitat-presence location; otherwise, it was considered a predicted habitat-absence location. To estimate the model accuracy, the percentage of predicted habitats among the observed habitats was calculated for each model and each testing datasets. This resulted in 168 percentage values.

3.3. Spatial comparison of habitat models

To generate habitat probability maps, the 24 habitat models were applied to the five environmental variables for the entire study area. This resulted in 24 habitat probability maps, each of which provided any location of the study with a probability predicting the possibility of the habitat occurrence. To simplify these maps, they were converted to binary maps predicting habitat-presence and habitat-absence locations. The percentage of habitat locations was calculated for each of these maps. Each of the 24 binary maps was compared against two other binary maps developed using the same habitat dataset and a different approach (e.g. the binary map based on LR and habitat data collected in February 2003 was compared against three maps: (1) the binary map based on ENFA and data collected in February 2003; (2) the binary map based on SLR and data collected in February 2003; (3) the binary map based on ANN & data collected in February 2003). In total, 24 pairs of models were compared against each other. For each pair, percent-age of pixels where two maps have the same prediction (i.e. both habitat-presence and habitat-absence locations) was calculated.

4. Results

4.1. Habitat–environment relationships

Three of the selected approaches have outputs (e.g. the *p*-values and coefficients) that quantify the habitat–environment relationships: ENFA, LR and SLR. We omit such outputs from ENFA, since they are published in our previous study (Li et al., 2009). Table 2 shows the coefficients and *p*-values of six LR models and six SLR models. In this study, if a variable has a *p*-value > 0.05, we consider that this variable is significantly related to mosquito habitats. Table 2 demonstrates that elevation, shown in Fig. 2, has the best predictive power among all variables as it is significant in all 12 models (it is always negatively related to the habitats). The heat-load index has the poorest predictive power, as it lacks statistical significance in all six time periods. In general, the results on the significance levels of variables for most of the LR and SLR models developed using the same training dataset are similar; it is the coefficients of variables that vary. For example, that the coefficient of heatload index is negative in LR model and is positive in SLR for data collected in May 2004. Despite of these differences, we found no obvious seasonal patterns in the *p*-values and coefficients.

4.2. Habitat model accuracy

As described in the previous section, our model validation procedure resulted in 168 percentage values. These percentage values are included in Table 3, which consists of three sub-tables. Each sub-table includes validation results for the models based on datasets collected in one of the sampling years. Each sub-table includes two parts: the left side describes validation results for the models based on February datasets, and the right side describes validation results for the models based on May datasets.

To determine if models based on habitat data from one season can be used to predict habitat locations at a different season, the prediction values of each method are summarized based on the seasons of testing datasets (values in grey cells in each sub-table). Based on sub-tables (a), (c) and (e), models based on data from dry seasons consistently perform about 5–15% better in predicting habitats in dry seasons in different years than predicting habitats in wet seasons in different years. Based on sub-tables (b), (d) and (f), models based on data from wet seasons also consistently perform 1–7% better in predicting habitats in dry seasons in different years than predicting habitats in wet seasons in different years. Results also show that models based on wet season habitats perform better in predicting wet-season habitats than models based on dry season habitats (Table 3).

The prediction accuracy based on training datasets is described in the third row of each sub-table. Based on these values, three of the approaches, LR, SLR and ANN can all correctly predict at least 76% of habitat locations. ENFA has a slightly poorer performance with prediction accuracies lower than 67%. The highest accuracy, 94.1%, is achieved by the SLR model based on the February 2005 dataset. Based on the same training dataset, four of the six ANN models outperform all other models.

The percentages of correctly predicted habitats from testing datasets in the same time periods are displayed in the fourth row of each sub-table. Based on these percentages, three of the approaches, LR, SLR and ANN, also produce satisfactory results, correctly predicting at least 75% of habitat locations. ENFA still has a relatively poor performance with prediction accuracies lower than 60%. The highest accuracy, 96%, again is achieved by the SLR model based on the February 2005 dataset. Based on the validation using the testing datasets, three of the six ANN models outperform other models and three of the six SLR models outperform other models developed using the same training dataset. By examining the percentages of correctly predicted habitats based on testing data in other time periods (the fourth to eighth rows of each sub-table), it is noticeable that more than half of these percentage values are slightly lower than the percentage values based on testing datasets in the same time period. However, all approaches can accurately predict over 76% of habitats in other time periods, except ENFA. When applied to testing datasets in other time periods, the frequency with which LR, SLR and ANN models outperform each other based on the training dataset in the same time period is similar. Only one of the six SLR models outperforms the LR and ANN models developed based on the same testing datasets. Among all models, the ANN model based on the May 2005 testing dataset achieves the highest level of prediction accuracy when applied to habitats observed in other areas.

4.3. Spatial comparison of habitat models

As shown in Table 4, the percentages of suitable habitat area in the study area predicted by 24 models range from 14% to 31%, On average, ENFA models predict fewer habitat locations compared with other models and ANN models predict fewer habitat locations compared with LR and SLR models.

Table 5 shows the percentages of pixels where any two models based on different modelling approaches but same habitat dataset have the same predictions. These percentages are used as an indicator showing the agreements between habitat models produced using different approaches. Table 5 shows that models based on different modelling methods have reasonable agreement with each other and in at least 84% of study area, they produce the same prediction results. Pairs of LR and SLR models in general are more similar to each other than other pairs of models.

5. Discussion and conclusions

5.1. The generality of spatially predictive habitat models

In this study, we investigated the temporal and spatial generality of four habitat modelling approaches. Our results provided answers to four critical questions that we asked of mosquito habitat modelling and management. The results from this study provide a promising answer to the first and second question – does environment–habitat relationships change over time and can habitat models based on data collected in a particular time be used to predict habitat occurrence in a different time? Consistent with our previous analysis using ENFA (Li et al., 2009), our results indicate that the habitat–environment relationship does change between different time periods. This indicates that it is better to use the data from the same season to predict habitat distribution in a particular season. Although we found no obvious seasonal patterns in the model parameters, our results indicate that dry season

habitats are easier to predict than wet season habitats and models based on wet season habitat data have good predicting power toward dry season habitats. If only one sampling of habitats is possible, wet season may be a better sampling time periods than the dry season. We also found that the changes in habitat–environment relationships have a limited impact on the accuracy of certain habitat models. Using data collected in one field survey in either the dry or rainy seasons, we have developed habitat models that can be used to predict habitats in other time periods, with satisfactory prediction accuracies. This has important implications for habitat management because habitat locations are relatively predictable and therefore localized habitat management could be considered.

Our third question: are there any benefits to producing pseudo-absence data for habitat modelling? Our results indicate that although ENFA models are entirely based on field observation, their accuracies are relatively low. A possible reason is that the difference between the environmental conditions of habitats and those of the entire areas is not significant enough for the ENFA models to sufficiently predict areas that are suitable for habitats. In contrast, the presence–absence models in this study achieved satisfactory results, providing evidence to support the use of random locations that are at least 100 m away from habitat-presence locations as pseudo-absence locations.

One further question addressed in this study pertains to the flexibility of modelling approaches and the impact of flexibility on the temporal generality of models. We found that the temporal generality of a LR approach is slightly better than an ANN approach based on the testing datasets in other time periods. This is probably due to the flexibility of ANN models, which may be overly fitted to the data including the noise in the data (e.g. errors in GPS data). Our results are consistent with other studies that compare a LR and an ANN approach (Dreiseitl and Ohno-Machado, 2002). The SLR approach is also more flexible than the LR approach, since it is capable of taking into account the spatial autocorrelation in dependent variables (Mertens and Lambin, 1997; Telford and Birks, 2005). Although our results indicate the existence of spatial autocorrelation in the habitat data (ρ is significant in all six SLR models in Table 2), the LR models on average outperformed the SLR models. Spatially, the predictions based on LR models and SLR models are similar. This further indicates that the flexibility of model approaches might not have a strong influence on the temporal generality of presence–absence models at least in the case of mosquito habitat modelling in African highlands. Our last question concerns the impact of spatial autocorrelation on the temporal generality of models. SLR was the only modelling approach that explicitly took into account the spatial autocorrelation of the data. We found its accuracy to be similar to the LR models. The possible explanation could be that the spatial structure in the dependent variable varies in different time periods. Our previous study indicates that the clustering level of habitats does indeed change across seasons and years (Li et al., 2009). Each of the SLR models is calibrated based on the quantification of the particular spatial structure of the dependent variable observed at a particular time period. This may limit its ability to generalise to other time periods.

5.2. Recommendations for mosquito habitat models

All four models used in this study have their own characteristics. ENFA does not require the data on the absence of habitats. Despite its moderate performance in predicting the locations of mosquito habitats, ENFA would remain as one of the options for habitat modelling if pseudo-absence data are proven to be unreliable. Based on training data of February 2003, the SLR model achieved the highest prediction accuracy among all 24 models, which means that a SLR approach could provide the most thorough understanding of the data. Based on training data, ANN models consistently outperformed all other models, except two SLR models. This also indicates that ANN can better identify habitat–environment relationships than LR and ENFA approaches.

However, when model generality is taken into account, the best modelling option for mosquito habitat modelling in African countries appears to be the simplest modelling approach: LR models. First of all, the models produced by LR have a good ability to generalise, as they can accurately estimate at least 75% of habitats using habitat data collected in a single field trip. Secondly, the LR models have a good agreement with more complex modelling approaches, namely, SLR models and ANN models. Finally, the ease with which they are constructed suggests that LR models remain the 'gold standard' until their performance is substantially improved upon.

5.3. Limitations and future directions

The use of larval habitat maps has become increasingly popular in malaria control programs (Coetzee et al., 2000). Our study adds to the confidence of focalized mosquito habitat management and provides simple solutions to habitat modelling. However, there are several limitations regarding the current study. First, many malaria endemic areas are lowland areas, which may have very different landscapes (Coetzee et al., 2000). Further investigations are needed to confirm that our modelling approach can be applied to other areas or other geographic scales. Second, we only collected habitat data twice a year and it is largely unknown whether habitat patterning can be extrapolated to other times of the year. Additional studies are needed to explore the spatial patterns of habitats in months other than February and May. Third, in using a threshold of 0.5, our calculations for predictive efficacy may lose important information. Rather than a binary outcome of presence versus absence, a per-habitat probability of occurrence may be more informative to program managers. Finally, as the emphasis of this study is on the generality of habitat models, many other aspects of the models were not explored here. SLR and ANN approaches may excel the LR method in other, unexplored ways. For example, the SLR approach can help reveal the underlying structure in the habitat data. The Bayesian approach for SLR model fitting can provide useful information on the stability of parameter estimation. The ANN models can be made more flexible by adding more hidden layers in the structure. Although we recommend the LR approach for mosquito habitat modelling, the other two approaches may yet prove better options if the modelling purposes require understanding the underlying structures of habitat data.

One of our recent studies revealed significant temporal variability in habitat locations and patterns (Li et al., 2009). Taking these two studies together, our results indicate that mosquito habitat locations may have great spatial and temporal variability within the areas suitable for habitats. Where observational studies are inappropriate for practical purposes, habitat models can be used to guide the targeting of habitats.

Acknowledgments

We thank Steve Munga, Emmanuel Mushinzimana and Andrew K. Githeko for their great efforts in organizing the field research and providing archived data. This work was supported by NIH Grant R01 AI050243. We would also like to thank the anonymous reviewers for their constructive suggestions and comments.

References

Akhwale WS, Lum J, Kaneko A, Eto H, Obonyo C, Bjorkman A, Kobayakawa T. Anemia and malaria at different altitudes in the western highlands of Kenya. Acta Trop. 2004; 91:167–175. [PubMed: 15234666]

Beven KJ, Kirkby MJ. A physically based variable contributing area model of basin hydrology Hydrol. Sci Bull. 1979; 24:43–69.

Bian L, Li L, Yan G. Combining global and local estimates for spatial distribution of mosquito larval habitats. Gisci Remote Sens. 2006; 43:128–141.

Boken, VK.; Cracknell, AP.; Heathcote, RL. Monitoring and Predicting Agricultural Drought: A Global Study. Oxford University Press; USA: 2005.

Brownstein JS, Rosen H, Purdy D, Miller JR, Merlino M, Mostashari F, Fish D. Spatial analysis of west nile virus: rapid risk assessment of an introduced vector-borne zoonosis. Vector Borne Zoonotic Dis. 2002; 2:157–164. [PubMed: 12737545]

Bulluck L, Fleishman E, Betrus C, Blair R. Spatial and temporal variations in species occurrence rate affect the accuracy of occurrence models. Global Ecol Biogeogr. 2006; 15:27–38.

Chubachi, R. Fourth Series, Biology. Vol. 37. Science Reports of Tohoku University; 1976. The efficiency of the dipper in sampling of mosquito larvae and pupae under different conditions; p. 145-149.

Coetzee M, Craig M, Sueur D. Distribution of African malaria mosquitoes belonging to the *anopheles gambiae* complex. Parasitol Today. 2000; 16:74–77. [PubMed: 10652493]

Defense Mapping Agency, Washington, D.C. Defense mapping agency production specifications for 1:50,000 scale topographic maps of foreign areas. 1980.

Dreiseitl S, Ohno-Machado L. Logistic regression and artificial neural network classification models: a methodology review. J Biomed Inform. 2002 Oct.(35):352–359. [PubMed: 12968784]

Gu W, Novak R. Habitat-based modeling of impacts of mosquito larval interventions on entomological inoculation rates, incidence, and prevalence of malaria. Am J Trop Med Hyg. 2005; 73:546–552. [PubMed: 16172479]

Guisan A, Zimmermann NE. Predictive habitat distribution models in ecology. Ecol Model. 2000; 135:147–186.

Hay SI, Snow RW, Rogers DJ. From predicting mosquito habitat to malaria seasons using remotely sensed data: practice, problems and perspectives. Parasitol Today. 1998; 14:306–313. [PubMed: 17040796]

Hirzel AH, Helfer V, Metral F. Assessing habitat-suitability models with a virtual species. Ecol Model. 2001; 145:111–121.

Hirzel AH, Husser J, Chessel D, Perrin N. Ecological-niche factor analysis: how to compute habitat-suitability maps without absence data? Ecology. 2002; 83:2027–2036.

Huettmann F, Linke J. Assessment of different link functions for modeling binary data to derive sound inferences and predictions. Computational Science and Its Applications—ICCSA. 2003:986–1986.

Killeen GF, Fillinger U, Knols GJ. Advantages of larval control for African malaria vectors: low mobility and behavioural responsiveness of immature mosquito stages allow high effective coverage. Malaria J. 2002a; 1:8.

Killeen GF, Fillinger U, Kiche I, Gouagna LC, Knols GJ. Eradication of Anopheles gambiae from Brazil: lessons for malaria control in Africa? Lancet Infect Dis. 2002b; 2:618–627. [PubMed: 12383612]

Lek S, Guagan JF. Artificial neural networks as a tool in ecological modelling, an introduction. Ecol Model. 1999; 120:65–73.

LeSage JP, Pace RK. Models for spatially dependent missing data. J Real Estate Financ Econ. 2004; 29:233–254.

LeSage, JP.; Pace, RK. Introduction to Spatial Econometrics. Chapman & Hall/CRC; 2009.

Li L, Bian L, Yakob L, Zhou G, Yan G. Temporal and spatial stability of anopheles gambiae larval habitat distribution in western kenya highlands. Int J Health Geogr. 2009; 8:70. [PubMed: 20021640]

Malakooti MA, Biomndo K, Shanks GD. Reemergence of epidemic malaria in the highlands of western Kenya. Emerg Infect Dis. 1998; 4:671. [PubMed: 9866748]

Mertens B, Lambin EF. Spatial modelling of deforestation in southern Cameroon: spatial disaggregation of diverse deforestation processes. App Geog. 1997; 17:143–162.

Minakawa N, Githure JI, Beier JC, Yan G. Anopheline mosquito survival strategies during the dry period in western Kenya. J Med Entomol. 2001; 38:388–392. [PubMed: 11372963]

Minakawa N, Seda P, Yan G. Influence of host and larval habitat distribution on the abundance of African malaria vectors in western Kenya. Am J Trop Med Hyg. 2002; 67:32. [PubMed: 12363061]

Morrison, ML.; Marcot, BG.; Mannan, RW. Wildlife–habitat relationships, concepts and applications. 2. The University of Wisconsin Press; Madison, WI. USA: 1998.

Recknagel F. Applications of machine learning to ecological modelling. Ecol Model. 2001; 146:303–310.

Rumelhart, DE.; Hinton, GE.; Williams, RJ. Storming Media. 1986. Learning Internal Representations by Error Propagation.

Shanks GD, Biomndo K, Hay SI, Snow RW. Changing patterns of clinical malaria since 1965 among a tea estate population located in the Kenyan highlands. Trans R Soc Trop Med Hyg. 2000; 94:253–255. [PubMed: 10974991]

Strauss B, Biedermann R. Evaluating temporal and spatial generality: how valid are species–habitat relationship models? Ecol Model. 2007; 204:104–114.

Telford R, Birks H. The secret assumption of transfer functions: problems with spatial autocorrelation in evaluating model performance. Quaternary Sci Rev. 2005; 24:2173–2179.

Utzinger J, Tozan Y, Singer BH. Efficacy and cost-effectiveness of environmental management for malaria control. Trop Med Int Health. 2001; 6:677–687. [PubMed: 11555434]

Yakob L, Yan G. Modeling the effects of integrating larval habitat source reduction and insecticide treated nets for malaria control. PLoS One. 2009; 4:491.

Yakob L, Yan G. A network population model of the dynamics and control of African malaria vectors. Trans R Soc Trop Med Hyg. 2010; 104:669–675. [PubMed: 20813387]

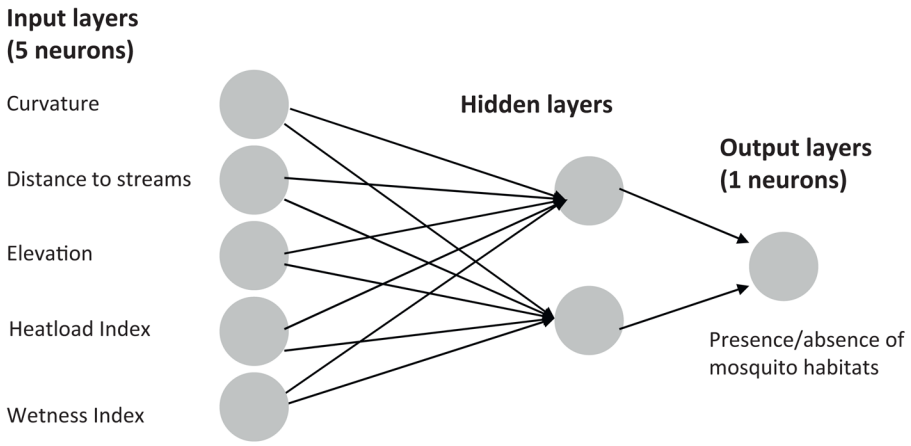

Fig. 1.
Illustration of a three-layered neural network with one input layer, one hidden layer and one output layer.

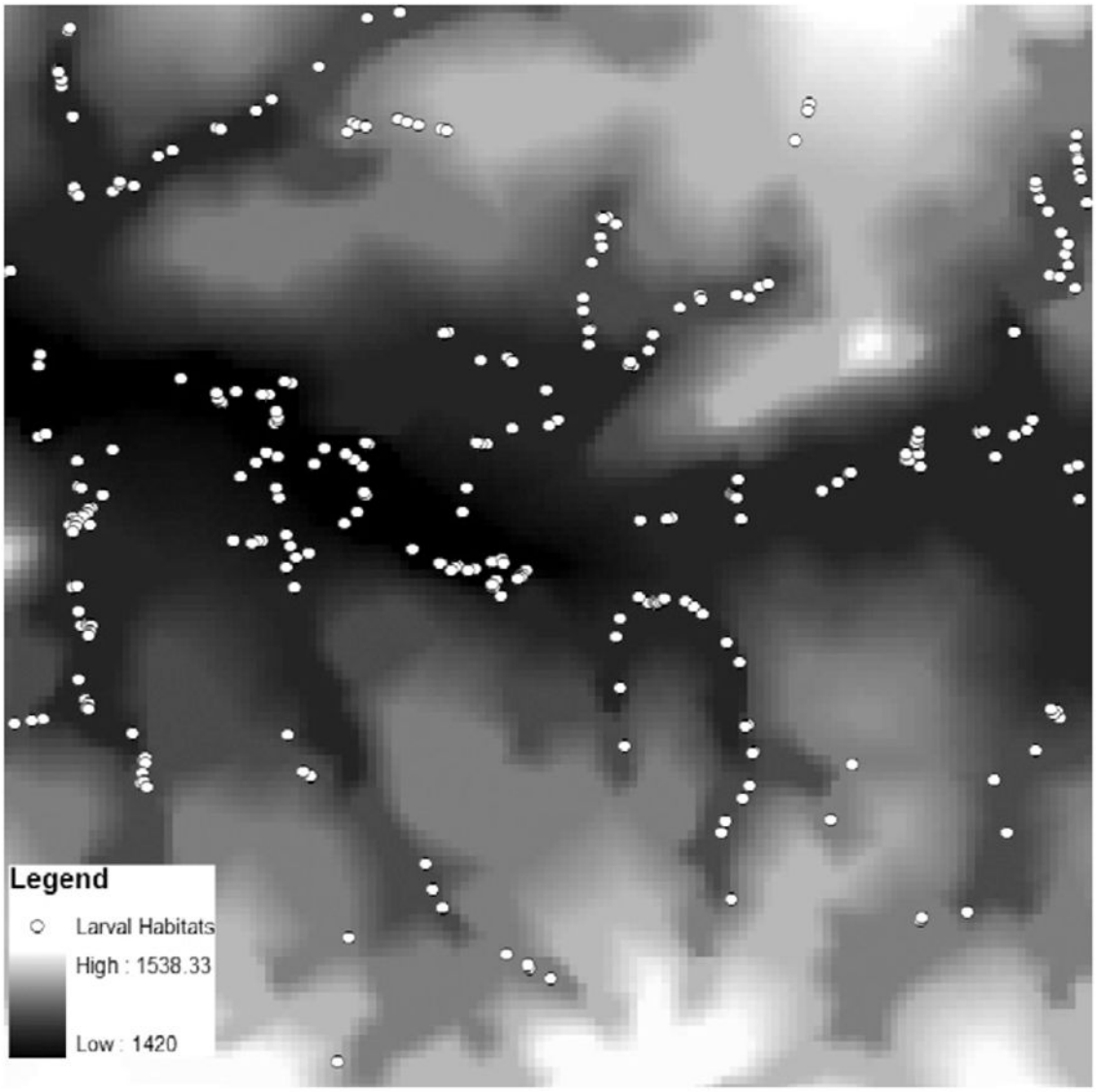

Fig. 2.
Map showing the elevation and distribution of larval habitats in May 2003 (white dots).

Table 1

Name, definition and ArcGIS extension for environmental variables used in the analysis.

Name	Definition	Value range	ArcGIS extension
Curvature	The probable degree of flow accumulation	−3.2 to 3.6	Spatial analyst
Digital elevation model	Elevation	1420–1540 m	Spatial analyst
Distance	Distance to the nearest streams	0–565 m	Spatial analyst & TAUDEM 4.0
Heat load index	The potential solar radiation (McCune and Keon, 2002)	1.0–1.3	Heat load index 1.0
Wetness index	Soil moisture level (Beven and Kirkby, 1979)	7.4–26.2	TAUDEM 4.0

Table 2

The parameters of LR and SLR models for mosquito habitats.

Approach	Models based on data collected in February 2003				Models based on data collected in May 2003			
	LR		SLR		LR		SLR	
Parameters	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**
Constant	37.7661	0.00	7.8472	0.09	48.5432	0.00	7.8310	0.02
Curvature	−0.0614	0.86	−0.1144	0.30	−0.4638	0.07	−0.5589	0.00
Distance to streams	−0.0073	0.00	−0.0033	0.00	−0.0061	0.00	−0.0024	0.00
Elevation	−0.0291	0.00	−0.0079	0.01	−0.0279	0.00	−0.0041	0.02
Heat load index	2.2470	0.69	2.0970	0.27	−7.1201	0.05	−1.9656	0.37
Wetness index	0.2045	0.00	0.1259	0.00	0.0604	0.04	0.0513	0.01
P	–	–	0.5723	0.00	–	–	0.7741	0.00

Approach	Models based on data collected in February 2004				Models based on data collected in May 2004			
	LR		SLR		LR		*SLR*	
Parameters	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**
Constant	40.3874	0.00	12.1259	0.08	59.8474	0.00	10.5306	0.04
Curvature	−1.1641	0.03	−0.8936	0.00	−0.3965	0.32	0.1352	0.44
Distance to streams	−0.0100	0.00	−0.0049	0.00	−0.0148	0.00	−0.0006	0.35
Elevation	−0.0224	0.00	−0.0057	0.03	−0.0326	0.00	−0.0088	0.00
Heat load index	−7.2965	0.26	−3.7757	0.40	−11.3682	0.04	0.6987	0.83
Wetness index	0.1000	0.11	0.0748	0.04	0.1198	0.01	0.1159	0.00
P	–	–	0.543	0.00	–	–	0.6719	0.00

Approach	Models based on data collected in February 2005				Models based on data collected in May 2005			
	LR		*SLR*		LR		*SLR*	
Parameters	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**	**Coefficient**	***p*-value**
Constant	77.4775	0.01	34.6773	0.00	57.83760	0.00	7.831	0.02
Curvature	1.1561	0.25	0.524	0.34	0.31690	0.36	−0.5589	0.00
Distance to streams	−0.0265	0.00	−0.0103	0.00	−0.0006	0.60	−0.0024	0.00
Elevation	−0.0439	0.01	−0.0170	0.02	−0.0394	0.00	−0.0041	0.02
Heat load index	−11.1048	0.42	−8.7731	0.19	−2.7809	0.60	−1.9656	0.37

Approach	Models based on data collected in February 2005				Models based on data collected in May 2005			
	LR		*SLR*		LR		*SLR*	
Parameters	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value	Coefficient	*p*-value
Wetness index	0.0723	0.52	0.0675	0.21	0.2069	0.00	0.0513	0.01
P	–	–	0.2284	0.18	–	–	0.7741	0.00

Table 3

Model accuracies.

	(1) Models developed using data collected in 2003								
	Models based on data in February 2003 (a)					Models based on data in May 2003 (b)			
Approach	**LR**	**SLR**	**ANN**	**ENFA**	**Approach**	**LR**	**SLR**	**ANN**	**ENFA**
Training dataset	86.5%	87.0%	87.0%	56.8%	Training dataset	82.7%	84.2%	83.3%	61.3%
Testing dataset	88.8%	82.7%	88.8%	42.8%	Testing dataset	79.5%	81.6%	80.3%	59.3%
May-03	79.1%	78.2%	78.2%	48.9%	February-03	87.4%	87.4%	88.2%	61.1%
February-04	81.6%	83.8%	81.2%	51.0%	February-04	87.6%	87.2%	86.5%	62.9%
May-04	84.7%	84.2%	87.0%	47.8%	May-04	90.2%	90.9%	89.1%	64.1%
February-05	91.1%	89.1%	95.1%	58.9%	February-05	96.0%	95.1%	95.1%	69.9%
May-05	78.5%	78.4%	71.9%	50.0%	May-05	81.1%	78.0%	81.5%	63.1%
Average of the above five testing results	83.0%	82.7%	82.7%	51.3%	Average	88.5%	87.7%	88.1%	64.2%
February average	86.3%	86.4%	88.1%	55.0%	February average	90.3%	89.9%	89.9%	64.6%
May average	80.8%	80.2%	79.0%	48.9%	May average	85.7%	84.5%	85.3%	63.6%

	(2) Models developed using data collected in 2004								
	Models based on data in February 2004 (c)					Models based on data collected in May 2004 (d)			
Approach	**LR**	**SLR**	**ANN**	**ENFA**	**Approach**	**LR**	**SLR**	**ANN**	**ENFA**
Training dataset	87.3%	86.6%	88.1%	56.7%	Training dataset	89.2%	92.1%	93.1%	58.5%
Testing dataset	84.9%	84.9%	88.0%	42.7%	Testing dataset	81.2%	82.6%	81.9%	40.5%
February-03	87.9%	86.4%	79.8%	46.9%	February-03	85.4%	86.9%	86.1%	51.5%
May-03	86.1%	85.7%	88.0%	48.7%	May-03	77.4%	79.1%	77.7%	49.3%
May-04	86.8%	87.0%	84.6%	48.5%	February-04	80.1%	84.6%	86.5%	53.6%
February-05	92.1%	95.1%	90.1%	57.5%	February-05	94.1%	93.1%	95.1%	63.0%
May-05	78.0%	77.6%	64.4%	52.5%	May-05	76.9%	76.3%	76.0%	53.9%
Average of the above five testing results	86.2%	86.4%	81.4%	50.8%	Average	82.8%	84.0%	84.3%	54.3%
February average	89.1%	90.4%	89.0%	53.1%	February Average	85.6%	86.4%	86.3%	54.6%
May average	84.2%	83.7%	76.3%	49.3%	May Average	78.5%	80.5%	81.2%	53.8%

	(3) Models developed using data collected in 2005								
	Models based on data collected in February 2005 (e)					Models based on data collected in May 2005 (f)			
Approach	**LR**	**SLR**	**ANN**	**ENFA**	**Approach**	**LR**	**SLR**	**ANN**	**ENFA**
Training dataset	92.2%	94.1%	88.2%	67.0%	Training dataset	75.8%	76.9%	77.7%	56.7%
Testing dataset	92.0%	96.0%	92.0%	59.0%	Testing dataset	75.0%	75.0%	75.0%	42.7%
February-03	87.9%	86.4%	79.8%	47.7%	February-03	87.6%	84.6%	87.4%	44.4%
May-03	86.1%	85.7%	88.0%	47.1%	May-03	78.2%	77.0%	79.0%	43.2%
February-04	86.1%	85.7%	88.0%	49.7%	February-04	80.1%	76.7%	85.7%	45.7%
May-04	86.8%	87.0%	84.6%	50.8%	May-04	88.6%	87.2%	91.2%	46.8%
May-05	78.0%	77.6%	64.4%	48.6%	February-05	95.1%	95.1%	98.0%	60.3%

	(3) Models developed using data collected in 2005								
	Models based on data collected in February 2005 (e)					Models based on data collected in May 2005 (f)			
Approach	**LR**	**SLR**	**ANN**	**ENFA**	**Approach**	**LR**	**SLR**	**ANN**	**ENFA**
Average of the above five testing results	85.0%	84.5%	81.0%	48.8%	Average	85.9%	84.1%	88.3%	48.1%
February average	86.4%	86.3%	86.3%	48.9%	February average	84.8%	82.9%	85.9%	44.8%
May average	84.0%	83.3%	77.4%	48.7%	May average	87.6%	85.9%	91.9%	53.0%

Table 4

Percentage areas that are predicted as habitats by 24 models.

	LR	SLR	ANN	ENFA
Models based on data in February 2003	26.7%	27.8%	26.2%	15.1%
Models based on data in May 2003	29.4%	31.7%	25.0%	21.0%
Models based on data in February 2004	28.8%	31.0%	23.2%	15.6%
Models based on data in May 2004	24.5%	26.6%	23.5%	15.1%
Models based on data in February 2005	24.9%	24.9%	29.3%	14.4%
Models based on data in May 2005	25.7%	23.8%	26.6%	14.1%
Average	26.7%	27.6%	25.6%	15.9%

Table 5

Agreements between models developed using different methods.

	LR vs. SLR	LR vs. ANN	SLR vs. ANN	ENFA vs. LR	ENFA vs. SLR	ENAF vs. ANN
Models based on data in February 2003	95.6%	89.2%	88.5%	88.0%	86.9%	84.6%
Models based on data in May 2003	92.2%	92.2%	91.1%	89.5%	87.4%	90.2%
Models based on data in February 2004	96.1%	89.0%	88.6%	85.7%	83.9%	85.0%
Models based on data in May 2004	95.8%	95.1%	94.1%	90.7%	89.9%	89.1%
Models based on data in February 2005	97.0%	90.0%	91.6%	85.9%	87.1%	84.6%
Models based on data in May 2005	91.9%	88.8%	88.7%	88.3%	89.8%	84.7%
Average	94.8%	90.7%	90.4%	88.0%	87.5%	86.4%