Zimbabwe Veterinary Journal 35(1):1-6, August 2017

ISSN: 1016-1511

Original Research

Comparison of GARP and Maxent in modelling the geographic distribution of *Bacillus anthracis* in Zimbabwe

Silvester M. Chikerema[1], Isaiah Gwitira[2], Amon Murwira[2], Davies M. Pfukenyi[1], and Gift Matope[3]

[1]*Department of Clinical Veterinary Studies, University of Zimbabwe, P.O. Box MP167, Mt Pleasant, Harare, Zimbabwe*

[2]*Department of Geography and Environmental Science, University of Zimbabwe, P.O. Box MP167, Mt Pleasant, Harare, Zimbabwe*

[3]*Department of Paraclinical Veterinary Studies, University of Zimbabwe, P.O. Box MP167, Mt Pleasant, Harare, Zimbabwe*

Correspondence: Silvester M. Chikerema, *schikerema@vet.uz.ac.zw*

Summary

A number of presence-only models can be used in the prediction of the geographic distribution of diseases and/or their vectors. The predictive performance of these models differs depending on a number of factors but primarily the modeled species' ecological traits. In this study, the performance of GARP and Maxent, two of the most commonly used modelling methods were compared in predicting presence and absence of anthrax in Zimbabwe using accuracy, sensitivity, specificity, Kappa statistic and the Jaccard coefficient as measures of model performance. The results showed that GARP had higher accuracy than Maxent (GARP = 0.70, Maxent = 0.67). Both methods had equal sensitivity (sensitivity = 0.71), but GARP had higher specificity (GARP=0.70, Maxent=0.67). Both Kappa and the Jaccard coefficient were also higher for GARP (0.335; 0.36) than for Maxent (0.295; 0.34). The results imply that GARP has superior performance over Maxent and is recommended for modelling species habitat suitability.

Keywords:
ENMs, GARP, Maxent, Anthrax

Received 10/03/2017. Accepted 21/07/2017.

Introduction

Habitat models, also known as ecological niche models (ENMs) are correlative models that use environmental and/or geographic information to explain observed patterns of species occurrence (Elith & Graham 2009). These models define the niche of the target species and predict its potential geographic and ecologic distribution through the analysis of relationships between combinations of environmental variables (i.e. bioclimatic, edaphic and topographical factors) and the species location data (Blackburn *et al.* 2007). The ecologic niche follows the Hutchinsonian definition as the hypervolume of ecologic parameters that allow a species to maintain populations without immigration (Peterson, Bauer & Mills 2004). ENMs have had wide applications in the study of biodiversity of flora and fauna, and recently have been used in epidemiology to understand disease ecology (Colacicco-Mayhugh, Masuoka & Grieco 2010; Gurgel-Goncalves, *et al.* 2012). There are a number of ENM techniques, ranging from those that require presence and absence data of the species under study to those that can use presence data only. The latter techniques include the Genetic Algorithm for Rule-set Prediction (GARP) (Stockwell 1999), and maximum entropy (Phillips, Anderson & Schapire 2006).

Surveillance and reporting of disease and/or disease vectors is a challenge to most developing countries like Zimbabwe due to limited resources. Consequently, complete data on spatial patterns of disease and disease vectors are limited. To overcome the problem of incomplete data, we can make use of the (few) presence records that are available, to predict areas where disease or disease

vectors are likely to occur (Peterson 2006). In this regard, predictive modelling provides a cost-effective alternative tool for estimating patterns of disease/vector distribution for informing public health authorities. Modelling ecological and environmental factors may give insights into which areas provide suitable habitat for vectors and pathogens (Peterson 2006). Predicted areas with a suitable habitat may then be taken to be at risk of the disease. This knowledge may then be used to improve surveillance and planning for future outbreaks. Accurate prediction of the potential spatial distribution of *B. anthracis* can be useful for generating new research hypotheses about disease persistence and for targeting surveillance efforts to areas at greater risk of potential disease presence, with the ultimate goal of providing sound methods for improving disease control. This is particularly important for resource-poor countries where disease surveillance and reporting capacity may be limited in spatial extent. Thus focusing the limited resources on areas at high risk may lead to better resource allocation.

However, the predictive performances of species distribution models (SDMs) may vary depending on the modeled species' ecological characteristics (McPherson & Jetz 2006).Thus, assessment of model performance is important in determining the suitability of the model for specific applications (Barry & Elith 2006). It also enables the user to investigate how different properties of the data and/ or of the species affect the accuracy of the maps generated by the techniques, and provides a basis for comparing algorithms (Segurado & Araujo 2004). In addition, it is important to use more than one metric to assess model performance because each quantifies a different aspect of predictive performance (Elith *et al*. 2006). Because of the importance of anthrax in Zimbabwe, which is a specified disease, and the need to accurately predict its spatial occurrence, this study sought to compare the performance of GARP and Maxent, two of the most commonly used modelling methods, in predicting the suitable habitat for *B. anthracis* using anthrax outbreaks data.

Materials and Methods

Anthrax occurrence data

The geographical locations of anthrax outbreaks were obtained from the Information and Management Unit (IMU) of the Division of Livestock Production and Veterinary Services in Harare, Zimbabwe. In this study, any confirmed anthrax occurrence was regarded as an outbreak regardless of the number of cases. A total of 110 geo-referenced records for the period 2005-2010 (for which bioclimatic parameters were available) of anthrax outbreaks were obtained from the IMU. After the data were converted from MGRS to Latitude and Longitude in Microsoft Excel spreadsheet it was then imported into a Geographic Information System (ILWIS 3.3 Academic) (www.itc.nl) for mapping.

Environmental variables

Climate data was freely downloaded from the global climate data site (www.worldclim.org/bioclim). To select variables that have an important influence on the distribution of anthrax outbreaks, the histogram method, as developed by Beaumont, Hughes & Poulsen (2005) was used. The histogram method shows the frequency distribution of values of each climatic variable throughout the species' range. According to the same authors, parameters with normally distributed and those with highly skewed values may have an important influence on a species' distribution and therefore should be included in modelling. Where there is no clear pattern in the histograms for a variable, the variable could be classified as irrelevant and therefore ignored for modelling. Similarly, where the histogram is normally distributed but is truncated in one or both tails, the parameter could also be rejected as this suggests that the species could tolerate other values of the parameter that were not included in the species' climatic envelope (Beaumont *et al*. 2005). After calculating histograms for the 19 bioclimatic variables from BIOCLIM, a jackknife test was carried out in Maxent (Philips *et al*. 2006), to determine variable importance and in this regard, variables with little contribution to the model were left out. The remaining variables were namely temperature seasonality, temperature annual range, mean temperature of the wettest quarter, precipitation seasonality and precipitation of the wettest quarter. The environmental variables were overlaid with the boundaries of Zimbabwe in a GIS and re-sampled to 90m spatial resolution to match the digital elevation model data for analysis.

Modelling techniques

The Genetic Algorithm for Rule-set Prediction (GARP)

GARP is a presence-only modelling technique that determines non-random associations between species occurrence localities (anthrax outbreaks) and environmental variables, such as satellite-derived data and interpolated field measurements (Stockwell 1999; Blackburn *et al.* 2007). Through an iterative process, GARP generates presence/absence predictions on the basis of a set of four logic rule types; i) atomic rules, where predicted locations are defined by a specific environmental variable; ii) range rules, where predicted locations are defined by a range of variables; iii) negated range rules, where predicted locations are defined as values outside of a defined range and; iv) logit rules, where predicted locations are fit to a logistic regression model with the environmental variables. The rules are developed through evolutionary refinement by testing and selecting rules on random draws of presence points from known occurrences data and pseudo-absences localities generated internally from the wider study area. In this study, we used Desktop GARP version 1.1.6 software freely available at www.lifemapper.org/*desktopgarp.*

Maximum Entropy (*Maxent*)

Maxent is a modelling technique which uses presence-only data to measure entropy, a measure of 'how much choice' is involved in the selection of an event (Phillips & Dudik 2004; Philips *et al.* 2006). Maxent is a general-purpose method for characterizing probability distributions from incomplete information. In estimating the probability distribution defining a species' distribution across a study area, Maxent formalizes the principle that the estimated distribution must agree with everything that is known (or inferred from the environmental conditions where the species has been observed) but should avoid making any assumptions that are not supported by the data (Philips *et al.* 2006). The approach is thus to find the probability distribution of maximum entropy (the distribution that is most spread-out, or closest to uniform) subject to constraints imposed by the information available regarding the observed distribution of the species and environmental conditions across the study area (Philips *et al.* 2006). The Maxent technique uses known occurrence locations (presence only data) and a set of gridded environmental layers to produce an output map of the predicted ecological niche of the species on a scale of 0 (lowest suitability) to 1 (highest suitability). Generally, if the area under the curve (AUC) is ≤ 0.5 the model is weak and no better than a random one, while the closer the AUC approaches 1 the better the prediction. In this study, we used Maxent version 3.3.3e, freely available at http://www.cs.princeton.edu/~schapire/Maxent

Model building

A total of 110 geo-referenced outbreaks were used as presence records. These were randomly split into 70% for training and 30% testing data initially run in Maxent. The 70% training data selected by Maxent were used for training the GARP model and the 30% retained for validating the models. A training/testing partition (70%/30%, respectively) internal to Desktop GARP was used for model building (Blackburn *et al.* 2007).

A total of 100 models were developed and the best subset procedure was employed to select the 20 best models under a 10% hard omission threshold and a 50% commission threshold fora final 10-model best subset (Blackburn *et al.* 2007). The final 10 models were summed within the GIS to visualize the geographic areas of presence/absence predicted across the best subsets. To enable model comparison, the testing data set was imported into a GIS system (DIVA-GIS) to generate pseudo-absence data. Dichotomous scores of presence-absence were generated after setting a presence threshold of ≥ 0.5 to produce maps of predicted disease presence/absence.

Evaluation and comparison of models

Model evaluation focused on the predictive performance of the modelling techniques and included the determination of a minimum threshold of quantitative output for the potential presence of the species. A 2x2 error matrix for each model was generated, and five measures of overall accuracy (rate of correctly classified pixels), sensitivity (probability of correctly detecting a presence), specificity (probability of correctly detecting an absence), Cohen's Kappa (overall accuracy corrected from that expected to occur by chance) and the Jaccard coefficient were calculated. Overall accuracy, defined as the rate of correctly classified cells, was used for each model.Two measures derived from the error matrix are sensitivity and specificity. Model sensitivity is defined as the proportion of true presences in relation to total presences predicted by the model (Allouche *et al.* 2006). Model specificity is the proportion of true absences in relation to absences predicted by the model (Allouche *et al.* 2006). The Kappa statistic normalizes the overall accuracy by the accuracy

that might have occurred by chance alone (Allouche *et al.* 2006). The Kappa statistic ranges from−1 to+1, where +1 indicates perfect agreement and values of zero or less indicate a performance no better than random (Cohen 1960).

Results

The two methods generated positive predictions in the central and north-eastern parts of the country (Figures 1a and b).The GARP had a higher overall accuracy, compared to Maxent (Table 1). The models had equal sensitivity, with specificity higher for GARP compared to Maxent. The Kappa value and the Jaccard coefficient for the two methods were in the same range of fair agreement, but higher for GARP compared to Maxent (Table 1).

Figure 1: Predicted presence/absence of anthrax by the **(a)** GARP and **(b)** Maxent models

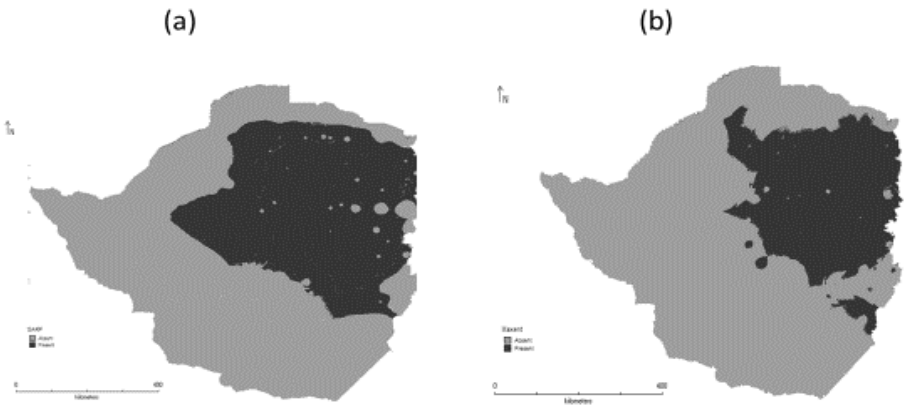

Table 1: Overall Accuracy, Sensitivity, Specificity, Kappa and the Jaccard Coefficient for the two models.

Modelling method	Overall Accuracy	Sensitivity	Specificity	Kappa statistic	Jaccard Coefficient
GARP	0.70	0.71	0.70	0.335	0.36
Maxent	0.67	0.71	0.67	0.295	0.34

Discussion

This study presents the first comparison of ENMs in Zimbabwe using anthrax occurrence data. The two methods predicted almost the same region in the central, east and north-eastern parts of the country, but with most performance metrics being higher for GARP than Maxent. Except for sensitivity, which was the same for the two methods, GARP showed relatively superior performance over Maxent in correctly predicting areas with and without anthrax (overall accuracy), and predicting areas without the disease (specificity).

The superior performance of GARP over Maxent is also supported by Anderson *et al*. (2003), who showed that GARP was superior over Maxent.

However, these results differ from those by Elith *et al*. (2006), and Elith & Graham (2009), who rated Maxent higher than GARP.

A number of factors may influence the relative performance of ENMs. Elith *et al*. (2006) observed that ENMs performance varied according to regions and the species studied. Regional variation was explained as due to differences in the quality of presence data available, with biases like collecting data from more accessible arears affecting model performance. Variation according to the species studied was explained as due to whether the species was a specialist or generalist, with specialist species being predicted better than a generalist, a result also observed by Segurado & Araujo (2004). While most performance measures used in this study were similar, the use of a minimum threshold (≥0.5) to enable calculation of performance measures limited comparison to the selected threshold (≥0.5), thus making performance comparison at lower threshold levels (≤) impossible.This could have masked differences at lower threshold since it has been shown some other measures like Kappa can vary at different thresholds (Larson *et al*. 2010).

Previous studies on anthrax in Zimbabwe (Chikerema *et al*. 2013) have shown soil type (a categorical variable) as an important variable in the spatial distribution of *B. anthracis*. Variable selection influences the estimated potential distribution of the species under study (Velez-Liendo, Strubbe & Matthysen 2013), thus in this case inclusion of soil type might have affected the predicted suitable areas for anthrax occurrence. However, according to Velez-Liendo, Strubbe & Matthysen (2013), although variable selection may influence the estimated potential distribution, it does not affect the relative performance of the model, which was the main objective of the present study.

According to Elith *et al*. (2006), the high predictive performance of machine learning methods like GARP and Maxent is due to their high level of flexibility in fitting complex responses, including the ability to handle interactions between the variables. It has also been suggested that the best models are those that minimise the omission error i.e. those that maximize sensitivity (Anderson *et al*. 2003), the reason being that incorrectly predicting a suitable habitat as unsuitable is a clear error, whereas predicting a suitable habitat where a species has not been observed may be due to insufficient sampling or other non-climatic factors limiting occupation by the species (Anderson *et al*. 2003). Using the performance measures employed in this study, the results show that the GARP has superior performance over Maxent and thus can be used for predicting the geographical distribution of the species, or other species of similar ecological traits. Thus, we make a claim that GARP is most suited for predicting the geographic distribution of *B. anthracis*.

Data Availability Statement

Data on the location of anthrax outbreaks is considered sensitive information by the Zimbabwean government as the disease is specified. For access to data, was granted by the Information and Management Unit (IMU) of the Division of Livestock Production and Veterinary Services in Harare, Zimbabwe.

Acknowledgements

The authors are grateful to the Division of Livestock Production and Veterinary Services Harare, Zimbabwe, for permission to access anthrax outbreaks data.

Conflict of interest

The authors declare that they had no financial or personal relationships that may have inappropriately influenced them in writing this manuscript.

References

Allouche, O., Tsoar, A. & Kadmon, R., 2006, 'Assessing the accuracy of species distribution models: prevalence, kappa and the true skill statistic (TSS)', *Journal of Applied Ecology* 43, 1223-1232.

Anderson, R.P., Lew, D. & Peterson, A.T., 2003, 'Evaluating models of species' geographic distributions: Criteria for selecting optimal models', *Ecological Modeling* 162, 211-232.

Barry, S. & Elith, J., 2006, 'Error and uncertainty in habitat models', *Journal of Applied Ecology* 43, 413-423.

Beaumont, L.J., Hughes, L. & Poulsen, M.P., 2005, 'Predicting species distributions: use of climatic parameters in BIOCLIM and its impact on predictions of species'current

and future distributions', *Ecological Modeling* 186, 250-269.

Blackburn, J.K., Mcnyset, K.M., Curtis, A.& Hugh-Jones, M.E., 2007,'Modeling thegeographic distribution of *Bacillus anthracis*, the causative agent of anthrax in the Contiguous United States using Predictive Ecologic Niche Modeling', *American Journal of Tropical Medicine and Hygiene* 77, 1103-1110.

Chikerema, S.M., Murwira, A., Matope, G. & Pfukenyi, D.M., 2013, 'Spatial modelling of *Bacillus anthracis* ecological niche in Zimbabwe', *Preventive Veterinary Medicine* 111(1-2), 25-30.

Cohen, J., 1960, 'A coefficient of agreement of nominal scales', *Educational and Psychological Measurement* 20, 37-46.

Colacicco-Mayhugh, M.G., Masuoka, P.M. & Grieco, J.P., 2010, 'Ecological niche model of*Phlebotomus alexandri* and *P. papatasi* (Diptera: Psychodidae) in the Middle East', *International Journal of Health Geographics,* 21 January, 9: 2.

Elith, J., Graham, C.H., Anderson, R.P., Dudik, M., Ferrier, S., Guisan, A., Hijmans, R.J.,Huettmann, F., Leathwick, J.R., Lehmann, A., Lucia, J.L., Lohmann, G., Loiselle, B.A., Manion, G., Moritz, C., Nakamura, M., Nakazawa, Y., Overton, M., Peterson, A.T., Philips, S.J., Richardson, K., Scachetti-Pereira, R., Schapire, R.E., Soberon, J.,Williams, S., Wisz, M.S. & Zimmermann, N.E., 2006, 'Novel methods improve prediction of species' distributions from occurrence data', *Ecography* 29, 129-151.

Elith, J. & Graham, C.H., 2009, 'Do they? How do they? WHY do they differ? On finding reasons for differing performances of species distribution models', *Ecography* 32, 66-77.

Gurgel-Goncalves, R., Galvao, C., Costa, J. & Peterson, T.A., 2012,'Geographical Distribution of Chagas Disease Vectors in Brazil Based on Ecological Niche Modeling', *Journal of Tropical Medicine* doi: 1155/2012/705326.

Larson, S.R., De Groote, J.P., Bartholomay, L.C. & Sugumaran, R., 2010, 'Ecological NicheModelling of Potential West Nile Virus Vector Mosquito Species in Iowa', *Journal ofInsect Science* 10(110), 1-17.

Mcpherson, J.M. & Jetz, W., 2006, 'Effects of species' ecology on the accuracy of distribution models', *Ecography* 30, 135-`151.

Peterson, A.T., Bauer, J.T., & Mills, J.N., 2004, 'Ecologic and geographic distribution of filovirus disease', *Emerging Infectious Diseases* 10, 40-47.

Peterson, A.T., 2006, 'Uses and requirements of ecological niche models and related distributional models', *Biodiversity Informatics* 3, 59-72.

Phillips, S.J., Anderson, R.P. & Schapire, R.E., 2006, 'Maximum Entropy Modeling of Species Geographic Distributions', *Ecological Modelling* 190, 231-259.

Phillips, S.J. & Dudik, M., 2004, 'A maximum entropy approach to species distribution modeling. *In 21st international Conference on Machine Learning*, Banff, Canada.

Segurado, P. & Araujo, M.B., 2004, 'An evaluation of methods for modelling species distributions', *Journal of Biogeography* 31, 1555-1568.

Stockwell, D.R.B., 1999, 'Genetic Algorithms II', In:Fielding, A.H., (ed). Machine learning methods for ecological applications', Kluwer Academic Publishers, Boston. pp 123-144.

Velez-Liendo, X., Strubbe, D. & Matthysen, E., 2013, 'Effects of variable selection on modelling habitat and potential distribution of the Andean bear in Bolivia', *Ursus* 24 (2), 127-138.