Ecological Niche Modeling of Zebra Species within Laikipia County, Kenya

Teddy Simon Mwangi, Hunja Waithaka, Mark Boitt

Department of Geomatic Engineering and GIS, Jomo Kenyatta University of Agriculture and Technology, Juja, Kenya
Email: tednetmail@gmail.com

Abstract
Wildlife conservation is essential, especially for countries like Kenya which rely on tourism as a major earner of foreign exchange. Conservation of species with minimal ecological information such as Grevy’s zebra, though a challenge, is critical to enable the future survival of such species. Grevy’s and Plains zebra have been classified as endangered and near-threatened by International Union for Conservation of Nature and Natural Resources (IUCN) respectively, with Grevy’s zebra found mostly in Northern Kenya and Ethiopia. This has been due to habitat degradation from livestock grazing, local hunting and development of resorts. Six prediction variables i.e. rainfall, temperature, land use, population, NDVI and cattle occurrence were used in Maxent algorithm to produce a habitat prediction map for both species. Both prediction maps had an AUC > 0.75, which is adequate for conservation planning. Niche similarity based on Warren’s I index (I = 0.78) indicates that both zebra species are identical based on their occupied niche environments, suggesting that similar conservation strategies can be adopted for both species.

Keywords
Grevy’s Zebra, Maxent, Niche Similarity

1. Introduction
Global tourism has experienced continued growth, accompanied with deepening diversification, to become one of the fastest growing economic sectors in the world, contributing 10.2% of global GDP in 2016. In Kenya, it contributed 9.8% of Kenya’s GDP in 2016, accounting for 9.2% of total employment in Kenya [1]. Tourism in Kenya is mainly wildlife-based, with the big five large mammals (elephant, rhino, buffalo, lion and leopard) being the main source of tourism satisfaction and revenue [2]. However, tourist’s preference on the big five has led
to the management policies being skewed towards their conservation, leading to under-appreciation of biodiversity. This tendency has led to heightened efforts to conserve the big-five, minimizing efforts to conserve other species [3]. Among the species under threat is the zebra species, of which Grevy’s and Plains zebra are the focus of this study.

Grevy’s zebra is a zebra species found in Kenya and Ethiopia. It has been assessed as endangered under criterion A2acd [4]. Grevy’s zebra population was approximately 2350 in Kenya in the year 2016 [4]. Factors identified that have led to declining population include habitat degradation, competition from livestock and local hunting [4]. Plains Zebra is listed as Near Threatened and it is close to qualifying for vulnerable under criterion A2a + 3c + 4ac [5]. Plains zebra population is estimated at 98,820 in 2016 [5]. Given the status of both zebra species, it is the focus of this study to see how similar they are based on their ecological niche.

Attempts to model both Plains and Grevy’s zebra niche have been few. Studies have majorly focused on enumeration of zebra population counts [6]. Kigen et al. modeled the potential current and future (2080) distribution of Grevy’s zebra in northern Kenya using a climate envelope model [7]. Kebede et al. modeled the seasonal habitat distribution of Grevy’s zebra in Alledeghi Wildlife Reserve (Ethiopia) with a view of developing conservation strategies [8]. However, inter-species niche comparison studies based on occupied geographic space have been few. One notable study focused on niche overlap of mountain hare subspecies in Europe vis a vis the invasion by the European hare [9].

In order to foster knowledge regarding the conservation of Grevy’s and Plains zebras and to provide insights that can be used for developing conservation efforts, this study has formulated two objectives: 1) To model and evaluate the ecological niche of both Grevy’s and Plains zebras and 2) To assess the level of niche similarity based on niche similarity indices.

2. Data and Methods

2.1. Study Area

Laikipia County is one of the 47 counties in the Republic of Kenya (Figure 1). It lies between latitudes 0°18’ and 0°51’ North and between longitude 36°11’ and 37°24’ East. It covers an area of 9462 km². The county mainly consists of a plateau bordered by the Great Rift Valley to the West, the Aberdare to the South and Mt. Kenya massifs to the South East. The annual average rainfall varies between 400 mm and 750 mm though higher annual rainfall totals are observed on the areas bordering the slopes of Mt. Kenya and the Aberdare Ranges. The county has a gazetted forest area of 580 Km² comprising of both the indigenous and plantation forests. The county population is estimated at 479,072 as of 2017 [10]. Laikipia county was chosen as the study area as it is where majority of the Grevy’s zebra are located in Kenya.
2.2. Data Collection

Data used in this study was classified into 3 categories: Environmental data (rainfall & temperature), Satellite Imagery (i.e. Landsat images to derive land use/land cover maps) & field survey data (zebra & cattle occurrence data). A summary of the data products used is shown in Table 1.

Rainfall and Temperature were adopted from Worldclim database, based on Kigen et al. (2003) work [7], where they modelled Grevy’s zebra in Northern Kenya using variables from Worldclim database. NDVI, LULC, cattle and population data acted as surrogate datasets, which represented the major causes that have affected the distribution and population of both zebra species [4]. NDVI represented the vegetation health hence areas where zebras could graze; LULC delineated the areas zebras are likely to be found based on known land cover and zebra occurrences; cattle represented the major competitor for zebras as they are both herbivores; and population represented the areas of human settlement.

Since the datasets had different spatial resolutions, all the variables were standardized to a resolution of 1 km using different resampling strategies. For
Table 1. Data used and respective data sources.

<table>
<thead>
<tr>
<th>Indicator</th>
<th>Data products</th>
<th>Source</th>
<th>Resolution</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rainfall</td>
<td>Worldclim</td>
<td>Worldclim</td>
<td>1 km</td>
</tr>
<tr>
<td>Temperature</td>
<td>Worldclim</td>
<td>Worldclim</td>
<td>1 km</td>
</tr>
<tr>
<td>Vegetation health</td>
<td>NDVI</td>
<td>Landsat</td>
<td>30 m</td>
</tr>
<tr>
<td>Land use/Land cover (LULC)</td>
<td>Classification</td>
<td>Landsat</td>
<td>30 m</td>
</tr>
<tr>
<td>Cattle occurrence data</td>
<td>Point data</td>
<td>Kenya wildlife service (KWS)</td>
<td>546 points</td>
</tr>
<tr>
<td>Grevy’s zebra occurrence</td>
<td>Point data</td>
<td>KWS</td>
<td>66 points</td>
</tr>
<tr>
<td>Plains zebra occurrence</td>
<td>Point data</td>
<td>KWS</td>
<td>594 points</td>
</tr>
<tr>
<td>Population</td>
<td>Population raster</td>
<td>AfriPOP</td>
<td>100 m</td>
</tr>
</tbody>
</table>

NDVI, an average window was used; for LULC, a majority window was used to depict the dominant land cover; and for population, a summation window was used to aggregate the total population within a 1 km grid.

2.3. Methodology Summary

A summarized workflow is shown in Figure 2. The first step is the interpolation of cattle data into cattle raster map. Predictor variables are then extracted at the zebra occurrence locations for each individual specie. The data is then split into a training and test dataset.

Testing or validation is required to assess the predictive performance of the model. Ideally an independent data set should be used for testing the model performance. However, in many cases this will not be available, a situation particularly prevalent in threatened and endangered species. Therefore, the most commonly used approach is to partition the data randomly into ‘training’ and “test” sets, thus creating quasi-independent data for model testing [11].

Variable selection is first done using multi-collinearity analysis on bioclimatic variables, and then jack-knife tests on the remaining variables. This was done so as to reduce model over-parameterization. Habitat distribution modeling was then undertaken based on the selected variables. What followed was an accuracy assessment that was done using the Receiver Operating Characteristics. If the result is acceptable, then niche similarity indices are computed. Detailed explanations of the analysis done are given in the subsequent sections.

2.4. Cattle Raster Interpolation

Cattle occurrence data was first interpolated using Inverse Distance Weighting (IDW) in order to develop a cattle occurrence raster map. IDW uses the measured values surrounding the prediction location to predict a value for the unmeasured location. IDW assumes that each measured point has a local influence that diminishes with distance. IDW makes the assumption that the value at the
unsampled location is the weighted average of known values within the neighborhood [12]. IDW formulas are outlines as

$$\hat{R}_p = \sum_{i=1}^{N} w_i R_i$$

where $\hat{R}_p$ means the unknown cattle data; $R_i$ means the cattle estimate at known locations; $N$ means the amount of known cattle locations; $w_i$ means the weighting of each cattle location; $d_i$ means the distance from each known cattle location to the unknown site; $\alpha$ means the power, and is also a control parameter, generally assumed as two [13]. Due to the difference in spatial resolution, all datasets were resampled to a resolution of 1 km so as to maintain spatial homogeneity.

### 2.5. Multi-Collinearity Analysis

Multi-collinearity refers to the pair-wise correlation among predictor variables based on the Pearson correlation coefficients. It affects the approximations of regression coefficients and induces bias responses between outputs and predictor variables [14].

Pearson correlation coefficients measure the strength and direction of the linear relationship between two variables, describing the direction and degree to which one variable is linearly related to another. It assumes that the variables are well approximated by normal distribution, and their joint distribution is bivariate normal [15]. It is denoted as follows

$$r_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$  \hspace{1cm} (2)
where $Y_{m,i}$ is the value of the measured inhibitory activity for compound \( i \) \((i = 1, 2, \ldots, 67)\), $Y_m$ is the average of the measured inhibitory activity, $Y_{est,i}$ is the value of the estimated inhibitory activity for compound \( i \), and $Y_{est}$ is the average of the estimated inhibitory activity.

The Pearson correlation coefficient can take values from $-1$ to $+1$. A value of $+1$ shows that the variables are perfectly linear related by an increasing relationship; a value of $-1$ shows that the variables are perfectly linear related by a decreasing relationship; and a value of 0 shows that the variables are not linearly related by each other. There is a strong correlation between variables if the correlation coefficient is greater than 0.8 and a weak correlation if the correlation coefficient is less than 0.5 [15].

Given multiple variables where the relationships between pairs of variables are many, a correlation coefficient matrix or a correlation ellipse can be developed. A correlation coefficient matrix is a matrix denoting the numerical pairwise correlation coefficient of each variable, organized in a matrix. A correlation ellipse plot is a visual plot that encodes the correlation coefficient into an ellipse, each having a different color code based on the strength of the correlation. The direction of the correlation (i.e. whether direct or inverse relationship) is also shown based on the direction of the semi-major axis [16]. Using correlation ellipses instead of correlation matrices allows for quicker understanding of underlying trends in the datasets.

2.6. Jack-Knife Tests

Jack-knife is an approach that excludes one variable at a time when running the model. In so doing, it provides information on the performance of each variable in the model in terms of how important each variable is at explaining the species distribution and how much unique information each variable provides. This can point out highly correlated variables, thereby allowing the user to determine if percent contribution values are likely to be skewed due to these correlations [17].

A model is created each time a variable is omitted from the model run in turn. Another model is also created using each variable alone. At the same time a model with all the variables in that Species Distribution Model (SDM) is also created. The Area under the Curve (AUC) of each model is recorded and all the values plotted together in the Jackknife. The Jack-knife tests thus shows the output of the AUC of the model with the following: All the variables; without one variable; and with only one individual variable in isolation. Comparing the 3 values gives an indication of the importance of each variable in predicting the species [18].

2.7. Habitat Distribution Modeling

Species Distribution Models (SDMs) have been employed widely in generating habitat distribution maps. Maxent is one of the Species Distribution Models that
has been applied widely in habitat distribution modelling. Maxent was chosen for this study because it is freely available, and has been proven to be one of the best performing modelling even with a relatively small number of samples [19] [20].

Maxent uses the principle of maximum entropy. Entropy is the measure of uncertainty associated with a random variable. The greater the entropy, the greater the uncertainty. Adhering to these concepts, Maxent utilizes presence-only points of occurrence, avoiding absence data and evading assumptions on the range of a given species [8]. It estimates the probability of occurrence of the species of interest using presence only species data and a group of environmental variables [21]. To obtain a solution, Maxent maximizes the gain function, a penalized maximum likelihood function. Exponentiating the gain function gives the likelihood ratio of an average presence to an average background point, hence maximizing the gain corresponds to finding a model that can best differentiate presences from background locations [22].

\[
\text{gain} = \frac{1}{m} \sum_{i=1}^{M} z(x_i) \lambda_i - \log \sum_{i=1}^{M} Q(x_i) e^{\sum_{j=1}^{J} \beta_j z_{ij}} - \sum_{j=1}^{J} \beta_j \sqrt{\frac{s^2[Z_j]}{M}}
\]

where the first term denotes sum of predicted values at presence locations, second term denotes sum of predicted values at background locations, and the third term an overfitting penalty, \(\beta\) is a regularization coefficient, and \(s^2[Z_j]\) is the variance of feature \(j\) at presence locations.

Maxent gives output data in the following formats: raw; cumulative; and logistic. The logistic format is recommended given that it provides estimates of the probability of occurrence as predicted by the included environmental variables [17].

### 2.8. Accuracy Assessment

The Area under Curve (AUC) curve is a widely used statistical technique for assessing accuracy of predictive models. An AUC plot is obtained by plotting the fraction of correctly classified cases on the y axis (sensitivity or true positive rate) against the fraction of wrongly classified cases (1-specificity or false positive rate) for all possible thresholds on the x axis at different threshold [23]. A random prediction will result in an AUC value of 0.5 whereas a perfect prediction assumes the maximum possible AUC of 1.0 [24]; AUC > 0.75 are considered as suitable for conservation planning [25]. AUC has been increasingly used in the evaluation of models of species distributions and is regarded as one of the best methods of assessments, with the primary advantage of providing a single measure of model performance that is independent of any particular choice of threshold [21]. The true and false positive rates are assessed at the species occurrence locations for each individual species.

It is important to note that Maxent computes a variation of AUC, based on the problem of classifying presence vs. background points (which may not be
true absences) [26]. However, since both zebra species are threatened to varying degrees and have dwindling numbers, the background points could be considered as pseudo-absences of species occurrence.

2.9. Niche Similarity

Conservation of species, both flora and fauna, has effects in both the ecological and evolutionary aspects. This can be assessed using SDMs. However, intra and inter-species interactions in ecological and geographical space is not as straight-forward, and methodologies for conducting niche similarity are not as developed as SDMs. Niche similarity is based on a test which asks whether SDMs from sister species predict one another’s known occurrence better than expected under the null hypothesis (that they provide absolutely no information about another’s range) [27]. Warren et al. (2008) proposed several niche similarity metrics that may be used to test niche similarity scenarios quantitatively.

Schoener (1968) statistic for niche overlap, ranges from 0 (niche models have no overlap) to 1 (niche models identical) [28].

\[ D(p_X, p_Y) = 1 - \frac{1}{2} \sum |p_{X,j} - p_{Y,j}| \] (4)

This metric is simple, has a long history of use, and permits direct comparison to traditional measures of niche similarity that focus on microhabitat and/or diet [27].

Another statistic used is based on Hellinger distances [29], defined as

\[ H(p_X, p_Y) = \sqrt{\sum (p_{X,j} - p_{Y,j})^2} \] (5)

Hellinger distances lie between 0 and 2. The similarity metrics have previously been used in ecological studies, primarily in comparing community composition across sites [30]. To compare Hellinger-based results to more conventional ecological measures of niche overlap, Warren et al. (2008) proposed a similarity statistic

\[ I(p_X, p_Y) = 1 - \frac{1}{2} H(p_X, p_Y) \] (6)

which ranges from 0 (no overlap) to 1 (niche models identical).

3. Results and Discussions

3.1. Correlation of Bioclimatic Variables

In order to reduce the potential of model over-parameterization, multi-collinearity analysis was conducted on the 19 bioclimatic variables. Variables with an inter-correlation higher than 0.8 (r > 0.8) were eliminated. To enable identify bioclimatic variables that have an inter-correlation higher than the set threshold, a correlation ellipse plot was developed, with the red dots identifying variables with a high level of pairwise correlation i.e. r > 0.8, hence one of the variables was eliminated. 10 of the 19 bioclimatic variables were found to be least corre-
lated (below the set threshold). The correlation ellipse plot was developed using
R programming for all the bioclimatic variables under consideration (See Figure
3).

3.2. Jack-Knife Tests

Jack knife tests were conducted after correlation, using 10 of the 19 bioclimatic
variables (which were least correlated) together with the following data sets: land
use/land cover; population; NDVI; and cattle occurrence raster. When the test is
run, variables contributing least to the model fit (<2%) were removed. The re-
mainning variables were then used for habitat distribution modelling for both
Grevy’s and Plains zebra.

Figure 4 shows the contribution of each variable to the overall model for the
Plains and Grevy’s zebra respectively. From the graphs, population, cattle, Bio19
and NDVI explain 77.5% of Plains zebra distribution. For Grevy’s zebra, cattle,
Bio4, Bio1 and population explain 80% of Grevy’s zebra distribution. Bio3,
Bio14, Bio1, Bio4, Bio13, Bio12 and Bio18 were thus eliminated for Plain’s zebra

![Figure 3. Correlation ellipse plot (red dots indicate r > 0.8).](image3)

![Figure 4. Variable importance for Plains and Grevy’s zebra respectively.](image4)
distribution modelling, while NDVI, Bio18, Bio14, Bio3, LULC, and Bio13 were eliminated for Grevy's distribution modelling.

3.3. Habitat Distribution Maps

Figure 5 shows the predictive probability occurrence maps of Plains and Grevy zebras based on the species occurrence and environmental variables. The logistic output was used so as to produce probability predictive maps, based on the maximized gain function.

The map shows that the core presence of both species is mainly in the middle section of the county, stretching from North to South. It indicates both species are located in the open woodlands and away from human habitats, though they face competition from cattle grazing. The concentration on the north is common to both species as these areas have a large concentration of wildlife conservancies and game reserves, hence human activity is regulated.

3.4. Accuracy Assessment

The habitat prediction maps were assessed based on area under curve (AUC) from a Receiver Operating Curve (ROC) plot. AUC > 0.75 was the threshold set for acceptance, as set out by Lobo et al. [15]. The ROC plots are a by-product of the Maxent software once it produces the habitat distribution maps (See Figure 6).

For Plains zebra, the AUC for training and test data was 0.82 and 0.79 respectively, while for Grevy’s zebra was 0.9 and 0.77 respectively. The jagged curves for Grevy’s zebra are as a result of few points available (66 compared to 594 for Plains zebra).

3.5. Niche Similarity

Schoener’s $D$ statistic and Warren’s $I$ statistic were both computed. Both showed significant levels of niche overlap as both $I, D \geq 0.7$ ($I = 0.782$ & $D = 0.706$).

Figure 5. Habitat distribution maps for Grevy’s and Plains zebra respectively.
indicating that the sister zebra species are identical based on their occupied niche environments. This suggests that similar conversation strategies can be adopted for both species.

4. Conclusion

This study conducts habitat distribution modelling and niche overlap, which is key for identifying if conservation efforts used for one species can be largely adopted in the conservation of another, especially endangered species. By considering bioclimatic variables, cattle data, land use and population metrics; a habitat distribution map of both Plains and Grevy’s zebra was developed; indicating concentration of both species around the centre of the county; running along a north-south direction to varying extents. This was validated using AUC plots, which indicated a training accuracy of 0.82 and 0.9 for Plains and Grevy’s zebra respectively. Niche similarity tests were then undertaken and found sig-
significant niche similarity based on the two metrics for the two species (i.e. $I = 0.782$ & $D = 0.706$). This indicates that Grevy’s zebra largely share the same ecological environment with the Plains zebra, which is significant in an effort to boost the endangered species numbers.

**Acknowledgements**

The author wishes to thank Kenya Wildlife Service for their assistance and cooperation while executing this study.

**References**


