DECISION TREE BASED CLASSIFICATION AND HABITAT SUITABILITY PREDICTION OF MIGRATORY BIRD: A CASE STUDY OF *Vanellus gregarious*

Vishal Kumar[1,2], Sameer Saran[1], C Jeganathan[2]
1Geoinformatics Department (GID), Indian Institute of Remote Sensing (IIRS),
Dehradun, Uttarakhand, India.
Email: vishalkumar@iirs.gov.in, sameer@iirs.gov.in
2Remote Sensing Department, BIT Mesra, Ranchi, Jharkhand,
Email: jeganathanc@bitmesra.ac.in

KEY WORDS: Machine Learning, Spatial Modeling, Bioclimatic variables

ABSTRACT: Habitat of any species can be defined as the biotic and abiotic components present in an area that supports the survival and reproduction of the species[15]. If these factors are altered due to any of the natural or anthropogenic reasons, the habitat becomes unsuitable for their survival. As a result of which, they either need to get adapted to the changes or find another suitable habitat. This is the key reason of migration observed in the migratory birds, i.e., when the conditions at their native sites becomes unfavorable for feeding, breeding or nesting, the birds migrate towards the regions with suitable climatic conditions. In this paper, the *Vanellus gregarious* (Socialbe lapwing) is considered for a case study, which is a winter migrant to India from Russia and Europe to analyze the prime environmental factors that constitutes the habitat of this species. The habitat is mainly influenced by ecological components like bioclimatic variables which includes temperature, precipitation and other climatic information, vegetation type or NDVI. The study is carried out by generating a set of decision rules considering the above components as parameters. The values from each component are used with the occurrences data of species from GBIF to derive rules. Further these set of rules are used as a knowledge classifier in decision tree for classification of suitable habitat. Decision tree is considered as one of the most user-friendly machine learning models because they make no linearity assumptions and automatically discover interactions among attributes. Each migratory birds have its own attributes as environment envelopes which provides different decision rules for classification. Such models can be used to analyze the spatio-temporal migration patterns. Also it can be helpful to understand species distribution, population dynamics and can guide in conservation or policy decisions for protecting threatened and endangered species.

1. INTRODUCTION

Habitat of a species is the environmental conditions that best support its growth, survival and reproduction[8]. If conditions in its habitat become unfavorable, they move to suitable places as seen in the case of migratory birds. Migratory birds travel thousands of miles for breeding or to escape the seasonal harsh conditions at their indigenous ground. *Vanellus gregarious* is one of the winter migrant that migrate India from Russia and Europe in winter season as winters in their native ground is unbearable for them. In this paper the example of *Vanellus gregarious* is considered for the study as it is listed as critically endangered (CR) species in the IUCN Red List of Threatened Species[5]. The population of this species have declined to 500 from 5000 in just one decade[14] and is moving rapidly towards extinction[5]. Indian migratory population of this species has become unseen with only few spots left for their sighting. The conditions pushing them towards extinction might be anthropogenic activities like conversion of long grasslands (here these wader nest) to agricultural lands or grazing [14] or the environmental factors like climate change. In this study the environmental variables and NDVI are considered as parameter to generate a set of rules to predict suitable habitat over the globe using machine learning algorithms. Machine learning techniques have gained rapid growth in past two decades. The major machine learning methods are supervised, unsupervised and reinforcement out of which the most commonly used is supervised learning[7]. Supervised learning performs complex computation by regression or classification techniques. This study deals with regression tree in context to the spatial and distribution data of species. The distribution data is occurrence of species at a particular location while the spatial data constitutes of

Fig. 1.1 - *Vanellus gregarius*
Source : IUCN, Jan-Michael Breider, www.pbase.com/breider

NDVI and 19 environmental variables. These variables affect species habitat and are changing gradually over a period of time. Due to highly modified environments, many difficulties have evolved in estimation of potential distribution of species [6]. This has raised the necessity to model habitat suitability of species with respect to biogeography, ecology, biology and climatic parameters.

2. STUDY AREA

The breeding and nesting area of *Vanellus gregarius* is mainly Russia and Kazakhstan but in winters they travel southern countries like Syria, Israel, India, Eretria and Sudan covering long distance via. Kyrgyzstan, Tajikistan, Uzbekistan, Turkmenistan, Afghanistan, Armenia, Iran, Iraq, Saudi Arabia and Turkey[16]. As the bird covers most of the central countries as shown in Fig. 2.1, so whole globe is taken as study area for this paper.

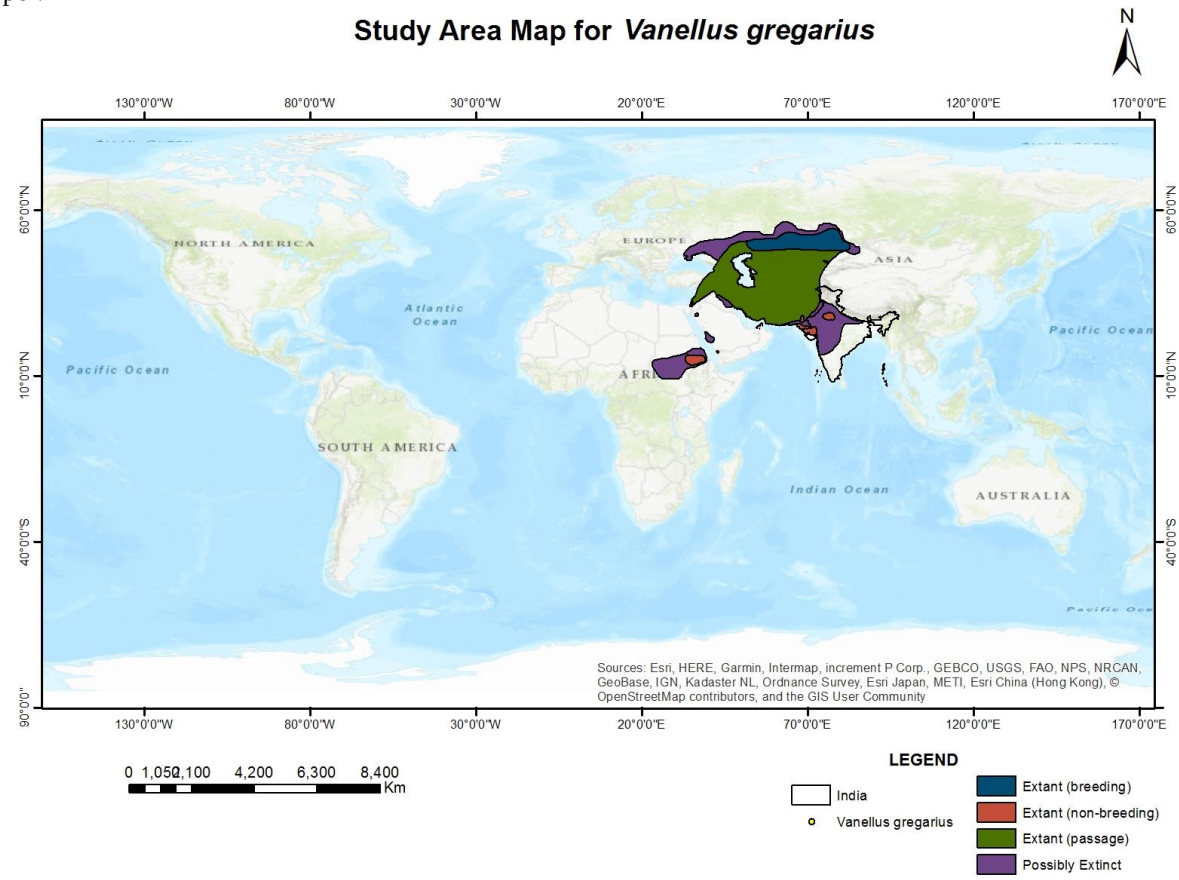

Fig. 2.1 : The Study Area map (Global) for *Vanellus gregarius*, Source: IUCN, for spatial extant.

3. MATERIALS and METHODS

3.1 Data Preparation

3.1.1 Species Data

The occurrence data of species is gathered by GBIF's online repository[10]. GIBF is a digital web-based repository on biodiversity containing occurrences, taxonomy and other relevant data of on wide range of species, which are collected and submitted by the globally recognized institutes/organizations, herbaria, museums, etc. It gives detailed data for every species, but for this study the other details were of less importance and are removed to make study focused data and enhance the model performance. The data were filtered by the record basis where the records made by human observations were extracted to avoid algorithm generated points. Further the records missing location information were excluded as the species data is later converted to spatial distribution map. The data is further filtered on the basis of recorded year to match with the year of spatial data.

3.1.2 Spatial Data

3.1.2.1 Environmental variables

The near current time (historical) global data of weather and climate is taken from WorldClim. The 19 bioclimatic variables version 2.1 is downloaded at a spatial resolution of 2.5 arc minutes (~4 km^2), a compiled data from 1970 to 2000. These bioclimatic variables are derived from monthly temperature and rainfall values to generate biologically more meaningful variables[1]. These variables represent annual, seasonal, quarterly, extreme or limiting trends of temperature and precipitation data. The 19 variables are shown in Fig. 3.1.

Fig. 3.1 : List of Bioclimatic Variables, Source: [1]

3.1.2.2 Vegetation Index (NDVI)

Vegetation composition and change are important factors that affect ecosystem condition and functions[9]. The global data for Normalized Differential Vegetation Index of MODIS monthly composite is downloaded from USGS, NASA website at the spatial resolution of 0.05 deg. (~5.6 km^2). This data is pre-processed and converted to 4km^2 to match spatial resolution of bioclimatic variables. Further the 19 bioclimatic variables and NDVI data is aligned in QGIS to same spatial reference in all terms like resolution, projection, extent etc.

3.2 Flowchart

The detailed flowchart followed throughout the study is displayed in Fig. 3.2. It consists of two types of data, spatial and non-spatial data. Non-spatial data constitutes species occurrence data which is later converted to spatial distribution map. The occurrence data is cleaned on basis of time, observations and location information, whereas spatial data pre-processing involves bringing all spatial data to identical spatial reference.

Data preprocessing and cleaning is an essential stage of any study as it involves removing errors and inconsistencies from data in order to improve the quality of data. Due to wide range of data inconsistencies, maintaining data quality is a vital task to carry any study[11]. After the data is cleaned and prepared, all the spatial layers are stacked to a single multiband raster image in R to make spatial data feasible to enhance model performance. At the same time the occurrence data is converted to distribution data of presence and pseudo-absence points are generated using R. Then these data are split to train and test data for model. Model is then trained and executed for prediction map. The model is then tested for the remaining data and the distribution points are overlaid on map which represents the same area as predicted by model.

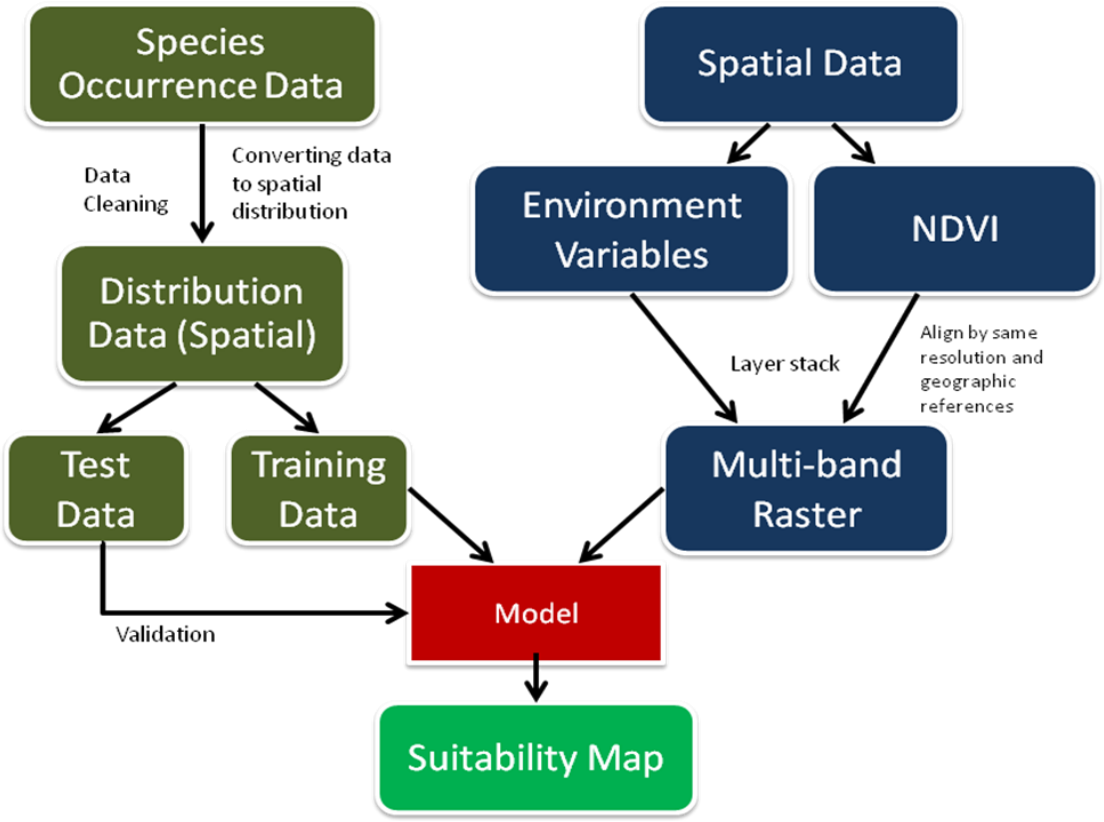

Fig. 3.2 : Flow chart for predicting habitat suitability of *Vanellus gregarius*

3.3 Habitat Suitability Modeling

Habitat suitability modeling has two focal points (i) environmental variables and (ii) interaction between variables and species[4]. There are ample existing machine learning models for species distribution modeling. Some of the most popular models used for habitat suitability modeling are maxent, random forest, bioclim, sdm, cart etc. This study uses gbm (Generalized Boosted Regression Model), it includes regression tree model which works on set of rules generated by using bioclimatic variables, NDVI and spatial distribution points to decide habitat suitability at a particular point.

The data of presence is taken from the pre-processed and cleaned occurrence data whereas the pseudo-absence points are generated in R, using the extent of presence data. The distribution data of species with presence and pseudo-absence points are then divided to training and test data. Now the spatial layers of bioclimatic variables and NDVI data are stacked together to generate a single multiband raster layer. Then the values of each variable are masked out at the distribution points (training points of presence and absence). After the values for each training points are extracted, these points act as nodes of decision tree with set of rules to predict habitat suitability for rest of the data. Now the model is trained with the training data and then, the GBM model is run over the whole globe to highlight suitable habitat areas and generate habitat suitability map.

4. RESULTS

In this paper the habitat suitability of *Vanellus gregarius* is predicted through GBM model on the basis of bioclimatic varialbes, NDVI and occurrence data using gbm (Genaralised Boosted Regression Model) model. This package is an extionsion to Freund and Schapire's AdaBoost algorithm and Friedman's gradient boosting machine. It includes various types of regression methods like least square, absolute loss, logistic etc[2],[12]. Further the gbm model is fitted by applying variuos settings in R. The gbm model automatically finds the optimal number of trees required for prediction based on the dataset provided. The model can be customized as per user requirement by providing additional informations as an argument to it[3]. The suitable habitat is then predicted using the trained model and suitability map is generated. The validation of model is then tested using test data which gave similar results as of trained model.

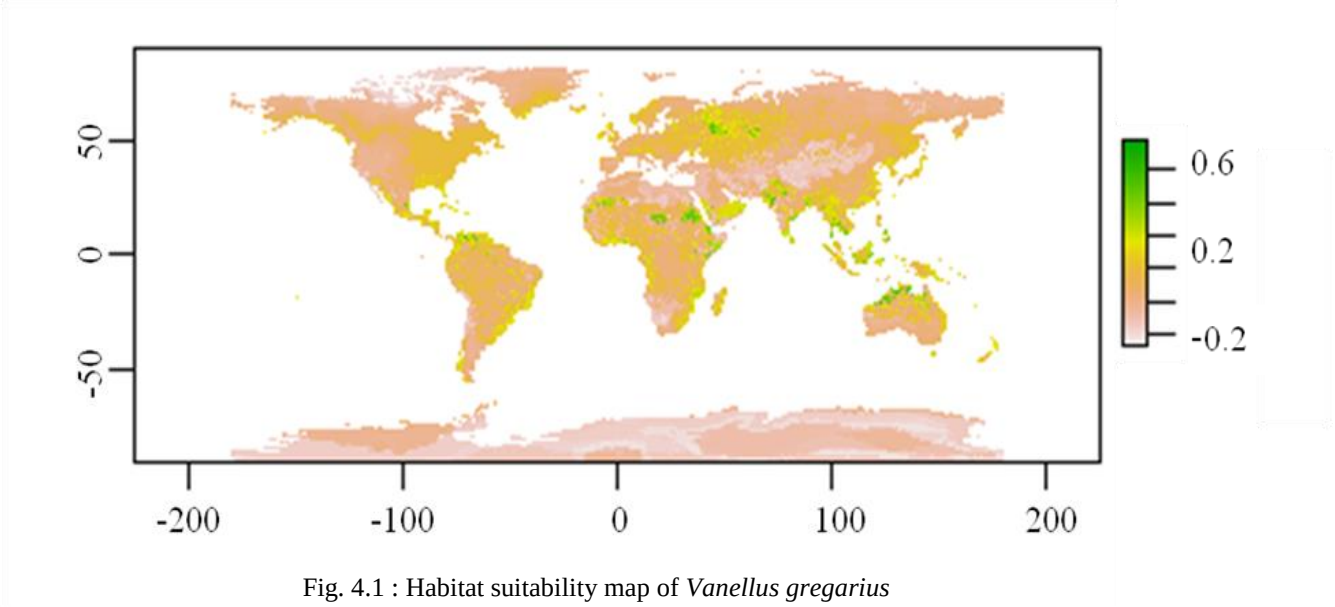

Fig. 4.1 : Habitat suitability map of *Vanellus gregarius*

The generated map in Fig.4.1 depicts the suitaiblity of Sociable lapwing. The green color in probability bar (right to the map) denotes high probability of suitable habitat while moving down the bar from green, yellow color denotes medium suitable condition, orange/brown represents low suitability, and white indicates no suitability. It is interpreted from the map that the suitable places includes north and eastern parts of Africa whereas in India, the north western part is found more suitable. Also, some parts of North and South America and Australia shows higer suitabbility in probability bar. The model is tested by overlaying distribution points on map, it is found that the distribution map prepared from occurrence data as shown in Fig. 4.1, represents similar distribution pattern as observed in downloaded occurrence data.

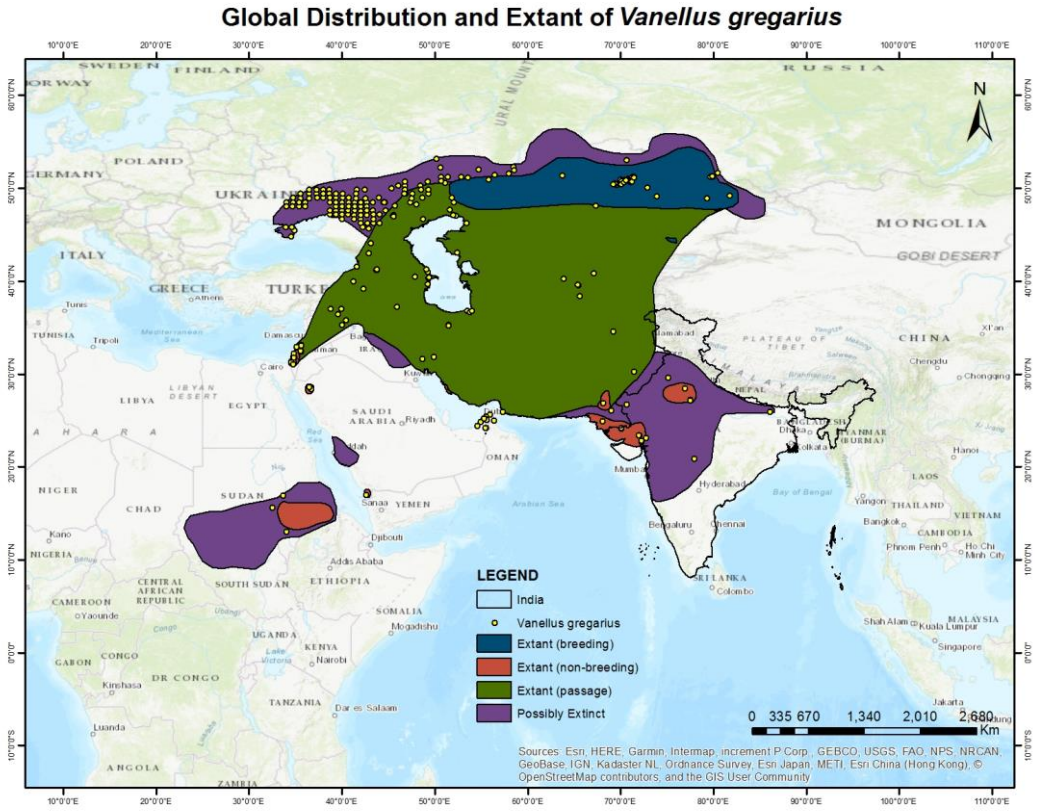

Fig. 4.2 : Distribution/occurrence points of *Vanellus gregarius* overlaid

5. DISCUSSION

There is an urgent need to take preventive measures for protection of migratory birds and other species moving towards the verge of extinction. As we all know, species in ecosystem are interlinked to each other, so if any one of species are removed from the system, the whole system gets misbalanced and the other species chained to that species will also be disturbed. There are various natural as well as anthropogenic activities directly or indirectly influencing the habitat of species. Their migration pattern and time is altered due to regular changing climate. By habitat suitability modeling we can identify and analyze the key climatic factors and possible locations which will be suitable for their future survival and make these sites as protected areas[13]. A major human hand in migratory bird's decline during migrations is their hunting. Various campaigns, programs or projects can be initiated to raise awareness among the people to protect these species.

REFERENCES

Alvarez, G., Salas, E.A.L., Harings, N.M. and Boykin, K.G., 2017. Projections of future suitable bioclimatic conditions of parthenogenetic whiptails. Climate, 5(2), p.34.

Breslow, N., 1972. Disussion of regression models and life-tables by Cox, DR. J. Roy. Statist. Assoc., B, 34, pp.216-217

Elith, J. and Leathwick, J., 2017. Boosted Regression Trees for ecological modeling. R Documentation. Available online: https://cran. r-project. org/web/packages/dismo/vignettes/brt. pdf (accessed on 12 June 2011).

Hirzel, A.H. and Le Lay, G., 2008. Habitat suitability modelling and niche theory. Journal of applied ecology, 45(5), pp.1372-1381.

IUCN Red List of Threatened Species. 2013. Retrieved 26 November 2013. https://www.iucnredlist.org/

Garzon, M.B., Blazek, R., Neteler, M., De Dios, R.S., Ollero, H.S. and Furlanello, C., 2006. Predicting habitat suitability with machine learning models: the potential area of Pinus sylvestris L. in the Iberian Peninsula. Ecological modelling, 197(3-4), pp.383-393.

Jordan, M.I. and Mitchell, T.M., 2015. Machine learning: Trends, perspectives, and prospects. Science, 349(6245), pp.255-260.

Kearney, M., 2006. Habitat, environment and niche: what are we modelling?. Oikos, 115(1), pp.186-191.

Lunetta, R.S., Knight, J.F., Ediriwickrema, J., Lyon, J.G. and Worthy, L.D., 2006. Land-cover change detection using multi-temporal MODIS NDVI data. Remote sensing of environment, 105(2), pp.142-154.

Orrell T, Informatics Office (2021). NMNH Extant Specimen Records. Version 1.45. National Museum of Natural History, Smithsonian Institution. Occurrence dataset https://doi.org/10.15468/hnhrg3 accessed via GBIF.org on 2021-09-19. https://www.gbif.org/occurrence/1322379838

Rahm, E. and Do, H.H., 2000. Data cleaning: Problems and current approaches. IEEE Data Eng. Bull., 23(4), pp.3-13.

Ridgeway, G. and Ridgeway, M.G., 2004. The gbm package. R Foundation for Statistical Computing, Vienna, Austria, 5(3).

Runge, C.A., Watson, J.E., Butchart, S.H., Hanson, J.O., Possingham, H.P. and Fuller, R.A., 2015. Protected areas and global conservation of migratory birds. Science, 350(6265), pp.1255-1258.

Watson, M., Wilson, J.M., Koshkin, M., Sherbakov, B., Karpov, F., Gavrilov, A., Schielzeth, H., Brombacher, M., Collar, N.J. and Cresswell, W., 2006. Nest survival and productivity of the critically endangered Sociable Lapwing Vanellus gregarius. Ibis, 148(3), pp.489-502.

Whittaker, R.H., Levin, S.A. and Root, R.B., 1973. Niche, habitat, and ecotope. The American Naturalist, 107(955), pp.321-338.

Wikipedia, BirdLife International (2013). "Vanellus gregarius". IUCN Red List of Threatened Species. 2013. Retrieved 26 November 2013. https://en.wikipedia.org/wiki/Sociable_lapwing