An analysis of propagated uncertainties in ecological niche spatial modelling

Marco Antonio Marinelli and **Robert Corner**
Department of Spatial Sciences
Curtin University of Technology
Perth, Western Australia
{m.marinelli, r.corner}@bom.gov.au

BIOCLIM is a probabilistic ecological niche model that can be used to investigate and predict species distributions for both native and agricultural species. Its results have greatest validity when studying relatively large (subcontinental) areas. The version of BIOCLIM used in this study uses three basic spatial climatic input layers (monthly maximum and minimum temperature and precipitation layers) and a dataset describing the current spatial distribution of the species of interest. Our work has investigated how uncertainty in the input data propagates through to the estimated spatial distribution for Field Peas *(Pisum sativum)* in the agriculturally significant region of south west Western Australia. Our results clearly show the effect of uncertainty in the input layers on the predicted specie's distribution map. In places the uncertainty significantly influences the final validity of the result and the spatial distribution of the validity also varies significantly.

Keywords: BIOCLIM; prediction; uncertainty;

I. INTRODUCTION

Spatial modelling, in the form of bioclimatic models is used to determine "environmental niches" in which species will thrive, both under present and predicted climates. Climatic factors and data concerning underlying biophysical variables such as soil properties, topography, drainage etc are significant drivers for these various spatial modelling tools. These inputs are typically in the form of raster surfaces, usually interpolated from sample data. Uncertainty, or error, in these input surfaces can propagate in a non linear fashion and radically affect the outcome of the predictions and recommendations made by the models (Burrough 1998). In this paper we investigate the BIOCLIM model (Nix 1986), which belongs to a subset of these models know as climate envelope models (Hijmans and Graham 2006). While these models ignore edaphic effects and consider climate as the sole driver for the suitability of a particular location for a particular species, we feel that this is justified since many edaphic factors are amenable to amelioration. Our study concentrates on the effect of the propagation of uncertainty through the modelling process rather than the validity of the model predictions per se.

The BIOCLIM model investigated has two groups of environmental inputs: (1) Twelve monthly interpolated temperature maximum, minimum and precipitation climate data layers (36 in total) and (2) a species distribution point map which shows where a species is known to survive. The initial step (in the overall process) is the calculation of 19 bioclimate layers, which further categorize seasonal trends and patterns (at each grid point). The next step inputs these 19 bioclimatic layers and the species distribution point map into the BIOCLIM model, which results in the final potential species distribution prediction.

A. The Environmental Envelope Method

The values of each of the 19 bioclimatic surfaces at each of the point data locations are used to develop the statistical distributions of sites where the species of interest in know to occur. BIOCLIM treats these as "multiple one-tailed percentile distributions, that is, it creates a percentile distribution for each variable so that, for example, the fifth percentile is treated the same as the 95th percentile" (Hijmans and Graham 2006). Then, on a grid cell by grid cell basis, all 19 layers are examined to determine the lowest percentile value out of the 19. This is regarded as the limiting variable and the appropriate percentile value is assigned to that grid cell as an estimate of output favourability. The output maps so produced therefore have a theoretical maximum for any cell of 50 and are scaled into the range 0-50. If the grid cell value for any of the variables falls outside the percentile distribution, a null value is allocated to the output for that cell.

B. Uncertainties in Input Climate Layers

Sources of uncertainty components in input surfaces arise in a number of ways. For current climatic variables, often used as baseline conditions, the uncertainty sources are measurement uncertainty, uncertainty due to long-term verses short-term averaging on a changing trend and uncertainty due to surface interpolation. For example, at the Perth Western Australia, rainfall station, the long term average of 1948–2007 is 790mm, whereas the decadal averages for 1948-57 and 1998-07 are 845mm and 738mm respectively. Maximum and minimum for this period are 1137 and 488mm respectively. Simply by using the differences in decadal averaging it could be argued that any calculation based on a long-term average and purporting to represent "Present conditions" for this station will have an uncertainty of 50mm.

The influence of two sources of uncertainty in the climate layers was studied. They are (a) uncertainties in the interpolated climate layers due to the interpolation method used and station density and (b) the uncertainty in these layers which results from the natural decadal variation in the

precipitation and temperature measurements that were used to generate these layers. Other sources of uncertainties and error have a smaller influence on the uncertainty in the prediction's accuracy and will not be further discussed.

II. Data Used

The interpolated climate layers used were the current climate Worldclim (2009) temperature maximum, temperature minimum and precipitation layers. These "current climate" layers represent the climate of the period 1950 to 2000. The main source of the precipitation and temperature data used to generate these surfaces was from the Global Homogeneous Climate Network dataset (NOAA 2008), which is a network of meteorological stations that have passed strict quality control criteria.

The crop studied was Field Peas (Pisum sativum) using, as the current species distribution, a dataset of successful field plot tests carried out across the study area.

III. Software Used

The BIOCLIM option, as bundled into AVID-GIS Dos (0.3), was used for the presence predictions (Hijmans et al. 2008). This was encapsulated in a purpose written IDL (6.4) program that allowed synthesised uncertainties to be added to the Worldclim climatic layers prior to calculating the 19 bioclimatic layers. The algorithms to generate these 19 layers, originally written in Arc Macro Language, were downloaded from the DIVA-GIS website and rewritten in IDL. ENVI (4.2) was used in the geospatial and statistical analysis of the simulated prediction results.

IV. Methods

The interpolated uncertainty in the Worlclim current climate layers are discussed and mapped in section 3.3 of Hijmans et al. (2005). From these relatively coarse resolution maps, the uncertainty in the precipitation and temperature, in the studied area, was estimated to be 2.5mm and 1.5deg C respectively for the entire studied region.

The decadal uncertainty layers were calculated from the GHCN point data. For each station, the difference in the decadal means (1950-59, to 1990-99) from the overall mean of that 50 year period was calculated. The largest absolute difference was then chosen as the maximum uncertainty (for each station) and from this, uncertainty surfaces were created by spline interpolation.

The Avid-GIS BIOCLIM option was executed from the IDL program. The model's inputs are the known species distribution and the Worldclim climate layers. The output is the predicted potential locations and their associated probability. The Monte Carlo method was used to investigate the sensitivity of the model to uncertainties in the climate inputs. The uncertainties were added to the Worldclim input layers. The maximum random uncertainty (for each grid point) was normally distributed and set to a maximum of the sum of the two uncertainty layers. The mean and standard deviation (S.D.) of these predictions (per grid location) did not stabilise until at least 4500 random realisations where generated. The mean represents the predicted output and the S.D. represents the propagated uncertainty. A set of predictions, referred to here as the default predictions, was also generated with no uncertainty. The ENVI Region of Interest tools were used to spatially analyse the effects of incorporating uncertainty.

V. Results

For the area studied, there are 23 discrete default BIOCLIM predictions. The values of the predictions, and the number of grid points where these occurred are shown in (Table 1). These predictions are in the percentile unit of measure defined for BIOCLIM and should not be confused with traditional statistical percentiles. The predictions have been grouped into 6 larger groups, with boundaries defined by a notably higher increase to the next prediction (approximately 2 percentiles). The Regions of Interest (ROI) are the grid points where the default predictions occurred. For example, ROI 2 is where the default prediction of 2.9 occurred. When the uncertainty layers was added to the climate inputs, the mean prediction at these grid points varies from 2.9, ranging from 0.74 to 9.49, with a mean of 2.89 and standard deviation of 1.11. These statistics will be referred to as the ROI-statistics.

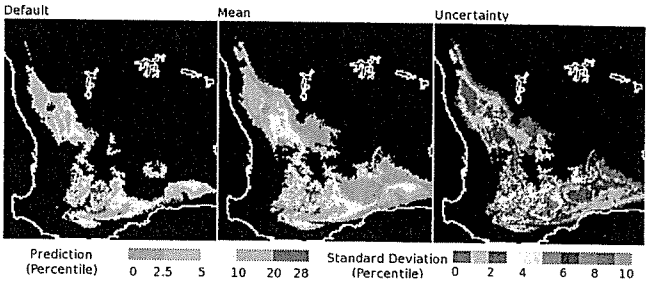

Figure 1. A BIOCLIM prediction, with and without uncertainty in input climate layers.

The default prediction, mean prediction and associated standard deviation at each grid point, is illustrated in Figure 1. The map clearly shows that the mean prediction can vary significantly from the default prediction and that the mean to S.D. relationship is not geographically consistent. The mean verses S.D. plot (see Figure 2) shows a complex change in the S.D. at each grid point (as the mean prediction increases at each grid point). To understand this relationship further, the grid points were analysed by ROI. For example, all coloured grid point in ROI 2.8 and 3.8 are shown in Figure 2, with these grid points further subdivided into 4 coloured sub regions:

1. Red - minimum mean prediction, to default prediction minus one ROI-S.D.

2. Green - default prediction minus one ROI-S.D., to default prediction.

3. Blue - default prediction, to default prediction plus one ROI-S.D.

4. Magenta - default prediction plus one ROI-S.D., to maximum prediction.

The ROI-S.D. of the predictions, within the examined ROI, was chosen as an arbitrary box within which most

measurements would lie. However, the centre of the study range is set to the default prediction (which defined the ROI).

TABLE I. REGIONS OF INTEREST.

ROI	Default Prediction (percentile)	Number of Points	Min	Max	Mean	SD
1	0	3126	0.01	5.56	0.66	0.79
2	2.9	867	0.74	9.49	2.82	1.11
3	3.8	1280	1.58	9.61	3.71	1.06
4	6.7	433	3.27	12.23	6.36	1.31
5	7.7	381	3.24	12.23	7.01	1.40
6	8.7	31	4.52	10.49	8.08	1.08
7	9.7	246	4.53	14.77	8.89	1.76
8	10.6	100	4.92	12.76	9.84	1.76
9	11.5	171	6.13	17.14	10.06	1.63
10	13.5	77	8.51	14.95	12.56	1.57
11	14.4	56	9.04	15.16	12.32	1.52
12	15.4	77	11.16	16.64	14.41	1.14
13	16.3	1	-	-	-	-
14	17.3	124	10.0	18.50	13.81	1.90
15	18.3	34	13.28	19.91	16.40	1.97
16	19.2	8	-	-	-	-
17	21.2	14	-	-	-	-
18	22.1	48	15.88	24.28	20.30	1.71
19	23.1	29	19.17	26.04	22.0	1.50
20	24.0	11	-	-	-	-
21	25.0	2	-	-	-	-
22	26.0	8	-	-	-	-
23	28.8	2	-	-	-	-

VI. DISCUSSION

For ROI 1 (in Group 1), the predictions are always greater than 0 (see Figure 3), so there are no red or green areas. For almost all of its grid points, the S.D. increases with increasing mean prediction. Relative to the mean prediction, the uncertainty is clearly greatest where the mean is lowest, before decreasing as the prediction increases. However, this ROI is where the S.D. is highest relative to the mean prediction (and BIOCLIM is most sensitive to uncertainties in the climate layers).

For all the ROI in Group 2 and most in Group 3, there is a "V" like relationship where the S.D. of the predictions (at a grid point) decreases as the mean of these predictions approaches their default prediction (red to green). Then, when the mean prediction passes the default prediction, the S.D. increases. This relationship is not evenly skewed, as the highest uncertainty occurs when the mean predictions are greater than one ROI-S.D. higher than the default prediction (the magenta coloured grid locations). For the ROI 5.6 and higher, the "V" like relationship is still visible, but to a lesser extent due to their being less grid points having higher mean predictions. Also, the direction of the skew is less clear and, in ROI 7.3 and higher, the S.D. is higher when the mean prediction is lower than the default prediction.

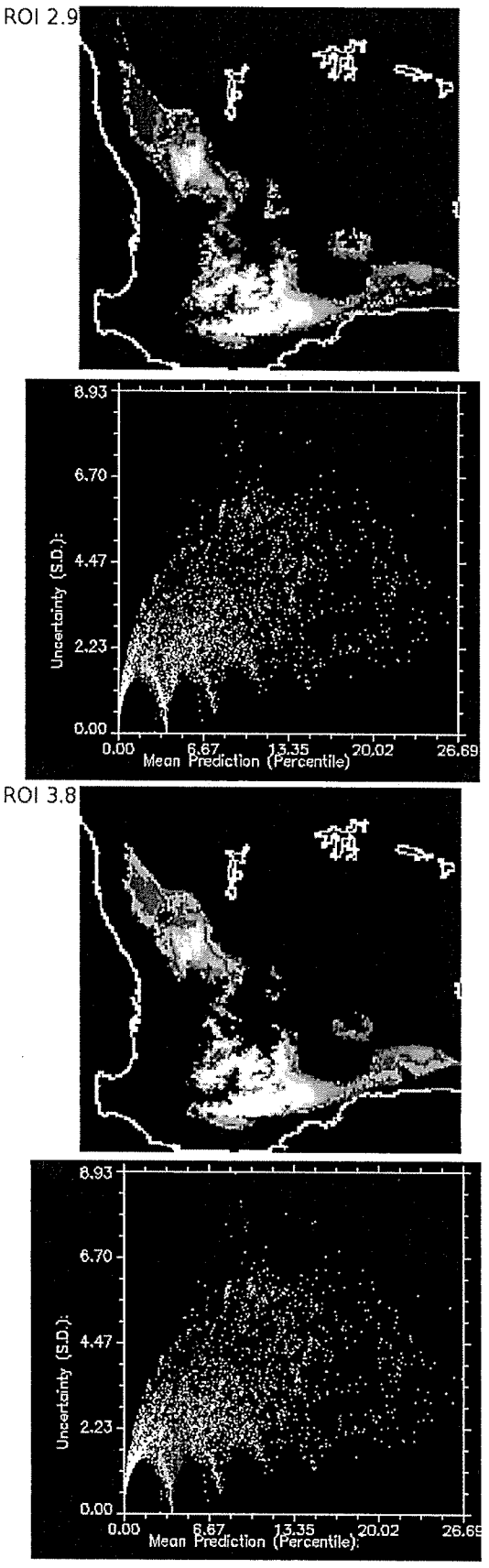

Figure 2. Uncertainty in the mean prediction. ROI 2.9 and 3.8.

The other significant pattern observed is that for some ROI (3.8, 7.7, 11.5 and 15.4), the majority of their grid points have lower uncertainties than observed in the other ROI in their Groups. This is most notable with ROI 3.8, which has grid points with a S.D. approaching 0, so the mean prediction (at these points) is clearly low when it is close to its related default value.

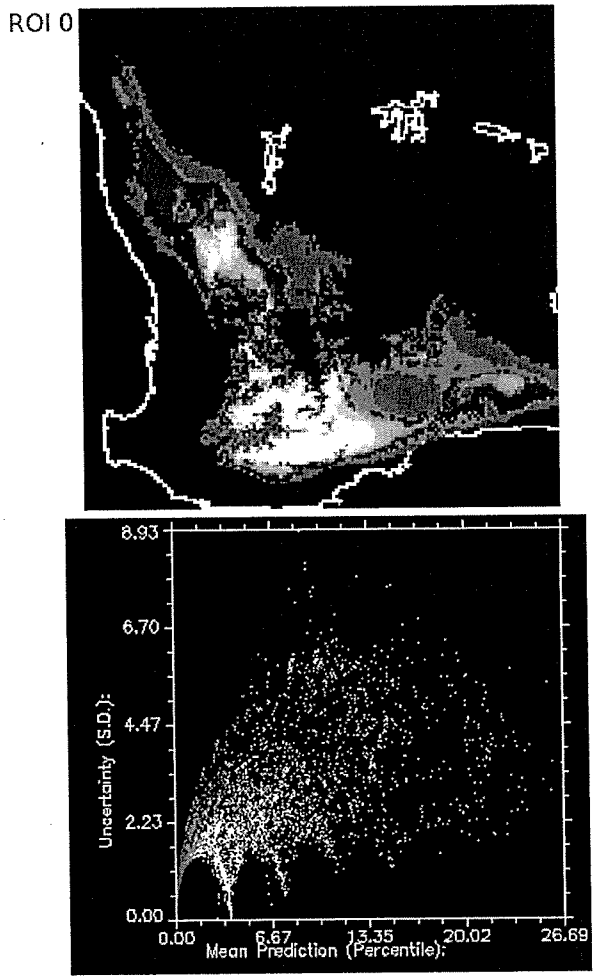

Figure 3. Uncertainty in mean prediction. ROI 0.

The spatial distribution of the grid point uncertainty varies significantly between each ROI. For example, In Group 1 (ROI 0) the most consistent pattern (in the S.D. of the predictions at each grid point, and in a large number of points), is along the north east border region. The range of predictions in these grid points centres on the lowest percentile range (above 0) that BIOCLIM can assign; which is 0-2.5. In contrast, the highest S.D. values (in these grid points) occur at the boundary of the region where the default predictions are in the 2.5-5.0th percentile range. This is consistent with the higher sensitivity-to-uncertainty observed at many grid points that border areas of different predictions. However, this is not a generalisation that can be applied to all of the ROI borders. For example, in the south west corner, where the default grid prediction quickly increases with changing location, from null to greater than 5th Percentile, the associated change and value in the calculated uncertainty is not significant.

VII. Conclusion

The uncertainty in the BIOCLIM prediction varies significantly, as does the prediction itself. However, the uncertainty does not have a clear relationship with the prediction, either spatially or with the predictions size. At some grid points, the uncertainty in the prediction can be very high relative to prediction, clearly showing the low validity of the predictions. On the other hand, at grid points relatively close to these, the uncertainty can drop significantly as the prediction increases. Ongoing work aims to help understand what, in the BIOCLIM model's structure, might cause this and other observed uncertainty to prediction relationships.

References

Burrough, P. A., McDonnell R. A. (1998). *Principles of geographic information systems*, 2nd Ed. Oxford: OUP.

Nix, H. (1986). A biogeographic analysis of Australian Elapid Snakes. *Atlas of elapid snakes of australia*. R. Longmore. Canberra, Australian Government Publishing Service: pp. 4-15.

Hijmans, R. J., S. E. Cameron, J. L. Parra, P. G. Jones and A. Jarvis (2005). Very high resolution interpolated climate surfaces for global land areas. *International Journal of Climatology*. 25, 1965-1978.

Hijmans, R. J. and C. H. Graham (2006). The ability of climate envelope models to predict the effect of climate change on species distributions. *Global Change Biology*. 12, 2227 - 2281.

Hijmans, R. J., A. Jarvis and R. A. O'Brien (2008). *DIVA-GIS v 5.4*.

NOAA. (2008). *Global Historical Climatology Network (GHCN-Monthly) data base* , from http://www.ncdc.noaa.gov/oa/climate/ghcn-monthly/index.php.

Worldclim – Global Climate Data (2009). http://www.worldclim.org/

IDL 6.4. (2007). *ITT Visual Information Systems*.

ENVI 4.2 (2007). *ITT Visual Information Systems*. http://www.itt.com/