Uncertainty in species distribution modeling – the use of Mahalanobis distance

Jon Olav Skøien[1], Gregoire Dubois[2], Gerard Heuvelink[3], Michael Schulz[4]

Abstract The Mahalanobis distance is a simple and commonly used method for species distribution modelling or niche modelling. Assuming that the range of a species depends on a set of environmental indicators, such as elevation, climate and land cover, we compute the mean vector and covariance matrix for these indicators at training locations, typically a protected area or locations where a species has been observed. The Mahalanobis distance can then be used as a similarity measure for these indicators for a larger region. Locations with high similarity can either be locations where it is likely to find the species, or locations which can be used for relocation of species when their current locations are threatened. Traditional niche modelling using the Mahalanobis distance does not take uncertainty of the input variables into account. In this paper, we discuss how to address uncertainty in the environmental indicators and how this influences the resulting similarity measure. In case uncertainties in input data are not documented, we present suggestions on how the spatial distribution (mean, variance, spatial correlation) of these uncertainties can be derived from the existing data. We analyze the propagation of uncertainty from environmental indicators to similarity measure both with analytical methods and a Monte Carlo approach. For the Monte Carlo approach we use sequential Gaussian cosimulations to generate realizations of the environmental indicators. Inclusion of uncertainty typically reduces the areas with high similarity and, at the same time, increases the areas with lower similarities. The approach discussed here was implemented in a Web Processing Service called eHabitat. While such a web based modelling service highlights the challenge of passing uncertain information in a web-based model environment (the Model Web), it also shows the advantages of having access to enhanced discovery tools, allowing the use of different data sets.

[1] European Commission Joint Research Centre, 21027 Ispra, Italy, jon.skoien@jrc.ec.europa.eu

[2] European Commission Joint Research Centre, 21027 Ispra, Italy, gregoire.dubois@jrc.ec.europa.eu

[3] Wageningen University, Land Dynamics Group, 6708 PB Wageningen, The Netherlands, Gerard.Heuvelink@wur.nl

[4] European Commission Joint Research Centre, 21027 Ispra, Italy, michael.schulz@jrc.ec.europa.eu

Ninth International Geostatistics Congress, Oslo, Norway June 11 – 15, 2012

Introduction

Species distribution models (SDMs) are typically used for prediction of the potential habitat of a species, based on observations of the species and a set of environmental indicators that are assumed to describe the niche of the species. There is a range of such models used in ecology [1]. A relatively common method is based on the Mahalanobis distance to create environmental suitability maps (ESM) [2-4]. Another method is the MaxEnt method [5] which is based on the creation of pseudo-absence locations.

Despite the common usage of these methods both for deriving species range maps and for ecological forecasting, little attention has been paid to the mathematical characteristics of the methods. Rotenberry et al. [4, 6] used principal component analysis to identify the variables that are most important for defining the niche of the species. Calenge et al. [7] recognized some limitations in the work of Rotenberry et al. [4, 6] and solved these by taking the available environmental space into account. Although based on a nice mathematical foundation and a clear improvement of Rotenberry et al. [4, 6], this method does not solve the basic problem of detecting whether the most suitable variables have been chosen. Variables with small variability within the reference area relative to the variability outside the reference area may lead to prediction of small suitable areas, however, the tolerance for changes in these variables (e.g. temperature) might be considerably larger than for some unmeasured variables (competition, access to the preferred food type, geographic limitations, etc).

We will not solve this problem in this paper either. However, we will show some theoretical results on predictions of habitat similarity, and in particular, analyze the uncertainty of the estimated suitability.

Methods

Mahalanobis Distance

For a set of environmental variables available for the region of interest, there are different ways of modelling the environmental similarity between this region and a reference geometry, typically a set of points referring to presence observations or points within a protected area where the species of interest occurs. The Mahalanobis distance D_i is used here as a measure of the similarity of a set of environmental variables between a pixel i and the averages of these environmental variables for the reference geometry, and is defined as:

$$D_{\mathrm{i}}^2 = (\mathbf{x}_{\cdot\mathrm{i}} - \mathbf{m})^{\mathrm{T}} \mathbf{C}^{-1} (\mathbf{x}_{\cdot\mathrm{i}} - \mathbf{m}) \quad (1)$$

where $\mathbf{x}_{\cdot i}$ is the vector of values of the environmental variables for pixel *i*, **m** is the mean of the environmental variables for the reference geometry, and $\boldsymbol{C}$ is the covariance matrix of the environmental variables for the reference geometry. The covariance matrix for *n* variables is given by

$$\mathbf{C} = \begin{bmatrix} \mathrm{cov}(\boldsymbol{x_1}\cdot, \boldsymbol{x_1}\cdot) & \mathrm{cov}(\boldsymbol{x_1}\cdot, \boldsymbol{x_2}\cdot) & \cdots & \cdots & \mathrm{cov}(\boldsymbol{x_1}\cdot, \boldsymbol{x_n}\cdot) \\ \mathrm{cov}(\boldsymbol{x_2}\cdot, \boldsymbol{x_1}\cdot) & \mathrm{cov}(\boldsymbol{x_2}\cdot, \boldsymbol{x_2}\cdot) & \cdots & \cdots & \mathrm{cov}(\boldsymbol{x_2}\cdot, \boldsymbol{x_n}\cdot) \\ \vdots & \vdots & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & \vdots \\ \mathrm{cov}(\boldsymbol{x_n}\cdot, \boldsymbol{x_1}\cdot) & \mathrm{cov}(\boldsymbol{x_n}\cdot, \boldsymbol{x_2}\cdot) & & & \mathrm{cov}(\boldsymbol{x_n}\cdot, \boldsymbol{x_n}\cdot) \end{bmatrix} \quad (2)$$

and the covariance between any two variables, $\boldsymbol{x_k}\cdot$ and $\boldsymbol{x_l}\cdot$, with means m_k and m_l and number of points in the reference geometry *J* is given by

$$\mathrm{cov}(\boldsymbol{x_k}\cdot, \boldsymbol{x_l}\cdot) = \sum_{j=1}^{J} \left(\frac{(x_{kj} - m_k)(x_{lj} - m_l)}{J} \right) \quad (3)$$

where x_{kj} and x_{lj} are the values of the indicators *k* and *l* at pixel *j*, respectively. The use of the inverse of the covariance matrix makes the Mahalanobis distance scale-independent, i.e., it is not affected by the different measurement scales of the variables. Also, highly correlated variables have less effect on D_i than uncorrelated variables. When the environmental variables used to generate the mean vector and covariance matrix are assumed normally distributed, then D_i^2 is distributed approximately according to a χ^2 distribution with *n*-1 degrees of freedom, and so we can convert D_i^2 into *p*-values. The *p*-values (or probability values) range from 0.0 representing no similarity to 1.0 for areas which are identical to the mean of the reference area. If the predictor variables are not normally distributed, the conversion is still useful as it rescales the unbounded D_i^2 values to a [0-1] range. As we generally do not assure normality of the input variables, we will in the following refer to this value as a similarity. Figure 1 illustrates the use of the Mahalanobis distance for identifying areas that have ecological characteristics similar to those found in a protected area, the Kafue National Park, in Zambia. A set of ecological variables are used as input data and a map of probabilities showing areas with similar and dissimilar values is generated.

One use of the predicted similarity is to identify areas with a similar combination of variables that could act as a replacement for the species in the park, should the conditions for some reason become unsuitable due to pressures. The idea would be that a park without similar areas outside the park must be protected more than one where there are suitable replacement areas. A simple index to summarize this is the habitat replaceability index (hri), which is the area outside the park with a similarity above a threshold, divided by the area of the PA. In our application we used a threshold of 0.5.

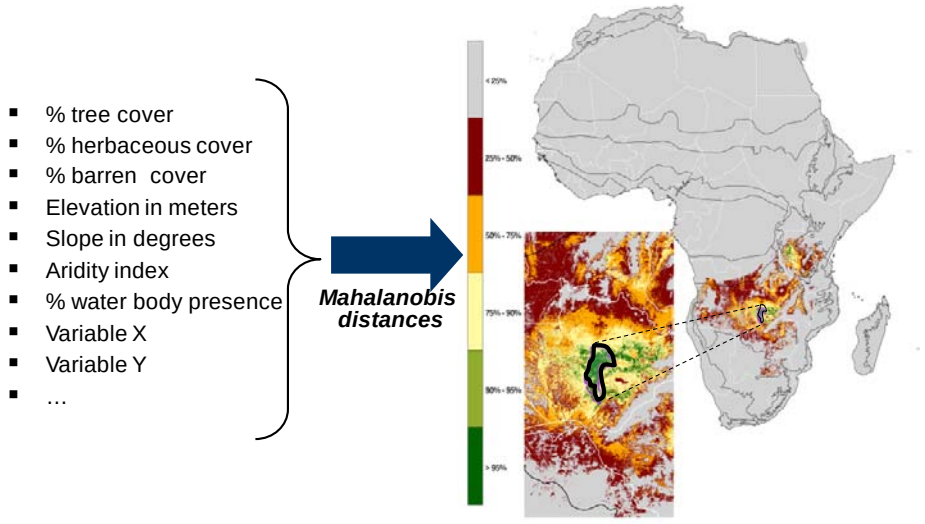

Figure 1 Map of habitats that are similar in the Zambezian ecoregion to the protected area of Kafue in Zambia, and associated scale of similarity.

Uncertainty of Mahalanobis Distance

The Mahalanobis distance for a pixel *i*, given in Equation (1) is a quadratic form, which can be generalized as (note that we for simplicity have not used the subscript *i* for the pixel to be predicted in the equations below):

$$D^2 = \mathbf{X}^{\mathbf{T}} \Lambda \mathbf{X} \quad (4)$$

where **X** is (**x-m**) and Λ is $\mathbf{C}^{-1}$. The expectation of this when **X** is a random vector with mean $\boldsymbol{\mu}$ and covariance matrix Σ is [8]:

$$E(D^2) = \mathrm{tr}(\Lambda\Sigma) + \boldsymbol{\mu}^{\mathrm{T}}\Lambda\boldsymbol{\mu} \quad (5)$$

where tr denotes the trace of the matrix (i.e. the sum along the diagonal). The variance of D is given by:

$$\mathrm{Var}(D^2) = 2\mathrm{tr}(\Lambda\Sigma\Lambda\Sigma) + 4\boldsymbol{\mu}^{\mathrm{T}}\Lambda\Sigma\Lambda\boldsymbol{\mu} \quad (6)$$

In the subsequent analysis we assume that **x** is random, not **m**. If we can assume stationarity, then the covariance matrix Σ of **x** at location *s* is composed of:

$$\Sigma_{ij} = \mathrm{Cov}(x_{is}, x_{js}) = \sigma_{is}\sigma_{js} - \gamma_{ij}(0) \quad (7)$$

where σ_{is} and σ_{js} are the local variances of the indicators and $\gamma_{ij}(0)$ is the cross-variogram between the two indicators at zero distance. The first term of Equation (5) always increases D^2 for increasing variances in Σ. This means that the expectation of D^2 increases hence and the expectation of similarity decreases if we assume that the values of the indicators are uncertain.

The above gets more complicated when also **m**, the means of the variables at the reference locations, are uncertain. In this case we are interested in the covariance matrix of **x-m**. The variance of indicator *i* is found as:

$$Var(x_i - m_i) = Var(x_i) + Var(m_i) - 2Cov(x_i m_i) \quad (8)$$

whereas the covariance between different indicators can be found as

$$Cov(x_i - m_i, x_j - m_j) = Cov(x_i, x_j) - Cov(x_i, m_j) - Cov(x_j, m_i) + Cov(m_i, m_j) \quad (9)$$

All variances and covariances involving the means can be found through integration of the cross-covariance between the individual points and all reference points, such as:

$$Var(m_i) = \frac{1}{J^2}[J + 2\sum_{k=1}^{J-1}\sum_{l=k+1}^{J} Cov_i(h_{kl})] \quad (10)$$

where $Cov_i(h_{kl})$ is the covariance function of variable *i* as a function of the separation distance h_{kl} between two different reference points *k* and *l*. The cross-covariance between one variable at prediction location *l* and the mean of another variable *j* can be found as:

$$Cov(x_i, m_j) = \overline{m_i x_j} + \frac{1}{J}\sum_{k=1}^{J} Cov_{ij}(h_{kl}) \quad (11)$$

where Cov_{ij} refers to the cross-covariance between two variables as function of separation distance h_{kl}. We do not give the remaining parts of the covariance matrices but these can be calculated in a similar way.

Simulation of uncertainties

The equations above are useful for interpretations, but they are not really feasible for computation of the Mahalanobis distance for a large number of points, typically up to a million points for a normal application. Instead in the following we show how simulations can be used for estimating the uncertainty. The approach we have chosen is to generate simulations of the input data, from a predefined distribution of the uncertainty. The simulations must reflect the spatial

correlation of the observed data, particularly if we are treating the reference variables as uncertain. Otherwise the variances in the covariance matrix inverted in Equation (1) might be too large, which would decrease the Mahalanobis distance.

Ideally, the simulations should be based on known uncertainty. However, many indicators come without any information about the associated uncertainty. The second issue is that indicators are usually cross-correlated, and hence their errors are also likely to be cross-correlated. This information is rarely available. We have therefore chosen to guesstimate the uncertainty in the following way:

- For each pixel and indicator, the standard deviation is estimated as 10% of the smallest value of the indicator above zero plus 5% of the indicator value at the pixel itself
- The correlations between the errors of the different indicators are assumed to be equal to the correlations between the different indicators
- A cross-correlogram of the uncertainties is based on the correlations and:
    - zero nugget effect (but easy to adapt for particular applications)
    - range found as the mean of the ranges of variograms automatically fitted to sample variograms of the individual indicators. Individually fitted ranges with higher value than the diagonal of the bounding box of the data set are set equal to the diagonal

From the cross-correlograms, we created sets of unconditional realizations of normalized errors of the indicators through sequential Gaussian simulation. The set of realizations is multiplied with the standard deviations given above, before adding the result to the initial data set. Realizations outside the ranges of the individual indicators are set equal to their minimum and maximum, respectively, to avoid impossible data, such as negative biotemperature or precipitation.

With a given number of simulations, we compute a set of possible maps of the similarities. From these we can either look at single realizations or different types of summary statistics. We present some possibilities below in the results section. All computations are done in R [9], the variogram fitting was done with automap [10], whereas the simulations were done with the package gstat [11].

Example data

As an example case study, we consider the effect of changing climate on the similarity of PAs. As environmental indicators, we computed the variables of Holdridge's life zones [12] based on data from WorldClim[5] [13], as these are

[5] http://www.worldclim.org

suitable for climatic classification (Figure 2). The life zones can be conceptualized with some variation of the variables. Here we use the following three variables:

- Mean annual precipitation (P)
- Biotemperature (annual average)
- Ratio of mean annual potential evapotranspiration (PET) to P: PETR

Biotemperature is the annually averaged temperature after replacing all temperatures below the freezing point with zero values, assuming that plants are dormant at lower temperatures. PET is obtained from the Thorntwaite equation [14], from the mean monthly temperatures and latitute. PETR is then found from PET/P, and is a dimensionless measure of the aridity. The values can be characterised from super-arid to super-humid, according to the left and bottom part of Figure 2. The Thorntwaite equation is one of the simplest means to compute PET, and is widely used in large scale computations.

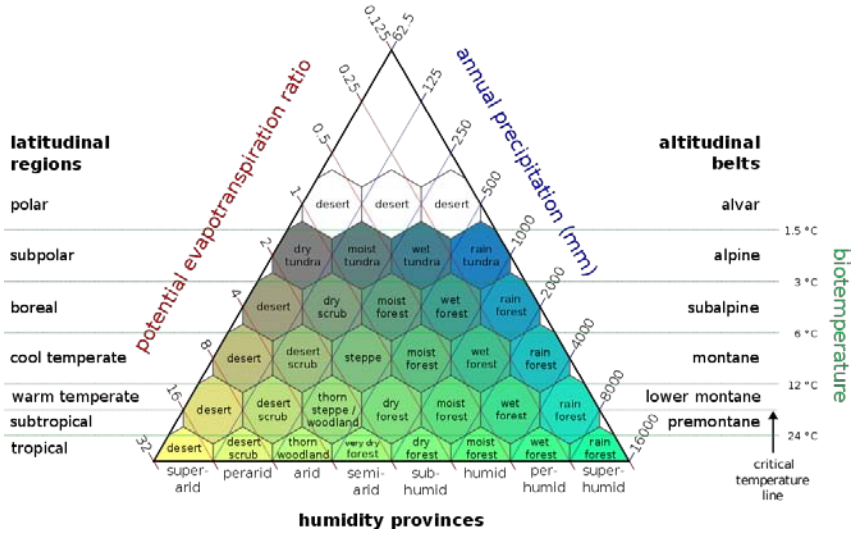

Figure 2 Holdridge's life zones based on different climatic indicators (Peter Halasz, original image published at: http://en.wikipedia.org/wiki/Image:Lifezones_Pengo.svg. Creative Commons Attribution Share)

Simulation of Mahalanobis distance

The Mahalanobis distance is a way of predicting the similarity between a region of interest (in our applications mainly a protected area) and the surroundings. Before presenting the effects of uncertainty on the Mahalanobis distance, we first present

some results where we predicted the similarity based on simulations. The setup is as following, all simulations with gstat [11]:

- Simulations on a grid with 200 cells in each direction
- Creation of a practically random correlation matrix for 10 variables, with cross-correlations in the range -0.54 to 0.99
- Creation of variograms and cross variograms based on the cross-correlations and a fixed range of 100 grid cells for a spherical model
- 10 unconditional simulations were created for each of the 10 variables
- A circular PA was defined at a random location, with a radius of 20 grid cells
- The similarity was for each set of simulations estimated for 2-10 variables

Figure 3 shows the similarities between the hypothetical park and the surroundings for an increasing number of variables, starting with 2 in the lower left corner and increasing to 10 in the upper right corner. Using just a few variables, we can notice that the similarity is high in many places outside the PA, but it is also quite low within the boundaries of the PA for a large number of pixels. With an increasing number of variables, the similarity decreases outside the park, whereas it increases inside the park.

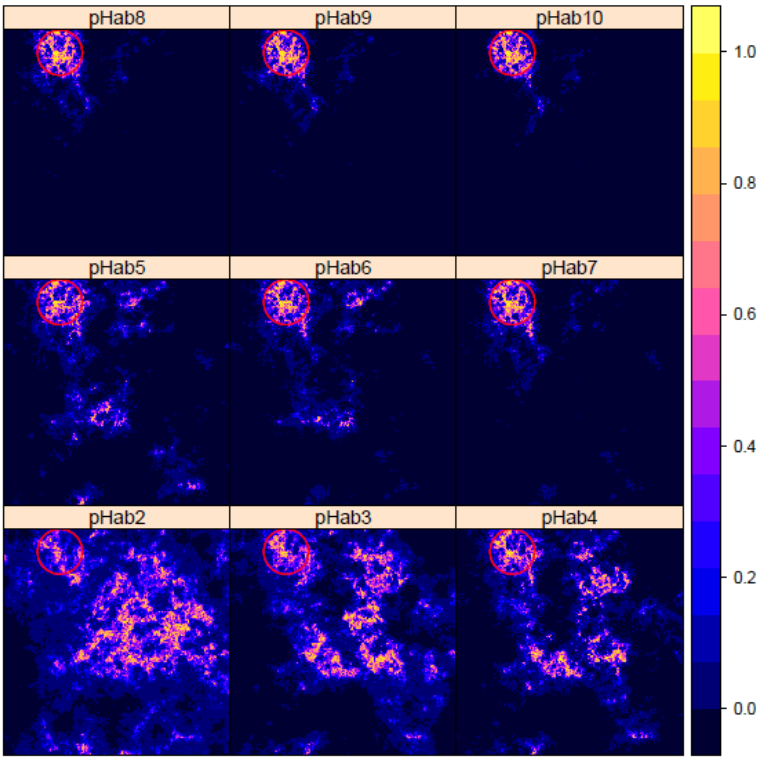

Figure 3 Similarities between a hypothetical circular PA and the surroundings for 2-10 variables (number in title)

Figure 4 shows the hri as a function of the number of variables for the 10 different sets of simulations. The figure shows that hri is generally decreasing with the number of variables for all simulations, although there are a few cases where an addition of a variable leads to a small increase of hri.

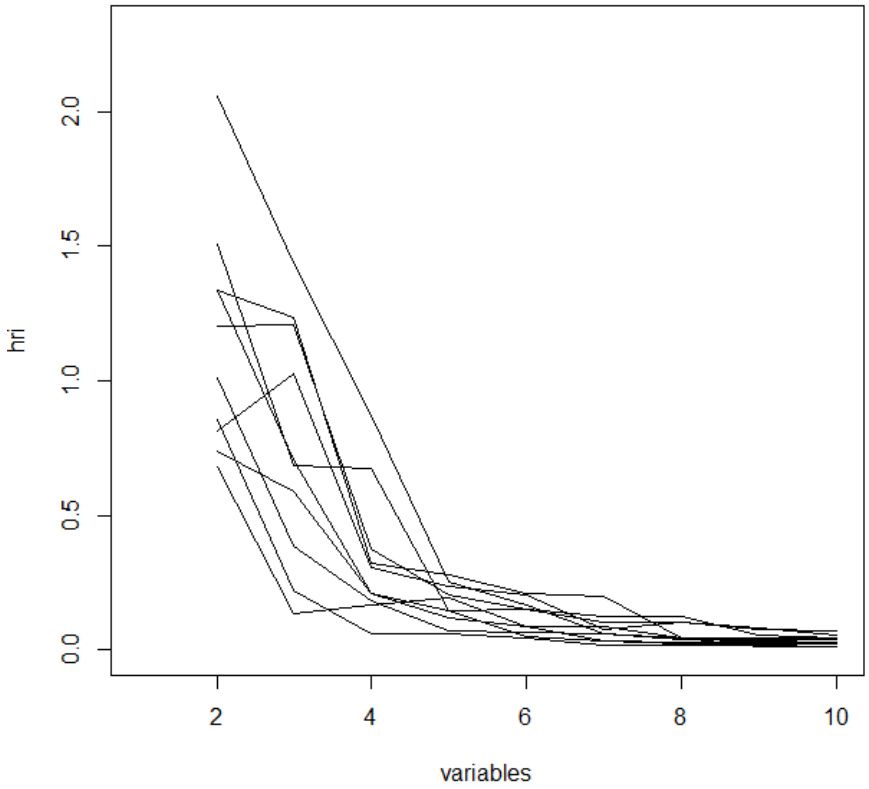

Figure 4 The hri as a function of numbers of variables for the 10 different sets of simulations

Results of uncertain species distribution modelling through a web service

The methodology above has been made available through the eHabitat Web Processing Service. More information about this service can be found at http://ehabitat.jrc.ec.europa.eu/ and in Dubois et al. [15]. We also have made the three variables described in the section above available as Web Coverage Services; these can be accessed through an example web client available at the web address above.. The user can choose a protected area (PA) of interest, current conditions for predictions of the similarity or future conditions for predicting the

future similarity to todays climate of the PA. The user can also choose the number of simulations to use in the predictions. We have used 25 in the example below.

Figures 5and 6 show the results from computing the bioclimatic similarities for a PA in the south-west of France (Landes de Gascogne). Both of these give a range of results in addition to the deterministic result, which is shown in the upper left panel in both figures. For the current situation, we can see that there is a high similarity around the PA itself, and also some areas in the north of Spain and Portugal. The two other panels in the top row show the results from two different sets of realisations of the input variables. Both suggest that there might be larger areas with some similarity than found from the deterministic computation, particularly around the PA. The second realisation gives lower similarity for the areas in Portugal and the north-west of Spain. The first panel to the left in the bottom row shows the mean of 25 simulations. The areas with a high similarity are almost the same as for the deterministic approach, but there are much larger areas also with lower similarity. The similarities are smoother than for the deterministic approach. The decrease in similarity that was suggested from Equation (5) can mainly be observed for the locations with the highest similarities. The second panel shows the standard deviation of the results. The map of the standard deviation is quite similar to that of the mean, as the standard deviation is higher for higher values. But there are some differences; the standard deviation is rather low inside the PA, and somewhat lower around the PA than in comparison to the areas with a high mean similarity in Portugal and the north-west of Spain. Thus, although the similarity is of similar magnitude, the uncertainty is much larger further away from the PA in this case, as could be expected from the spatial correlation that we introduced in the simulations. The last panel to the right shows the probability of similarity of a pixel to be larger than 0.5. This exceedance probability is useful to find areas with a high similarity independent on the uncertainty, and is high for the pixels inside and around the PA, but lower for the other areas that also have a relatively high similarity for the deterministic result and the mean from the simulations. This is again most likely a result of the spatial correlation of the simulated errors based on the uncertainty.

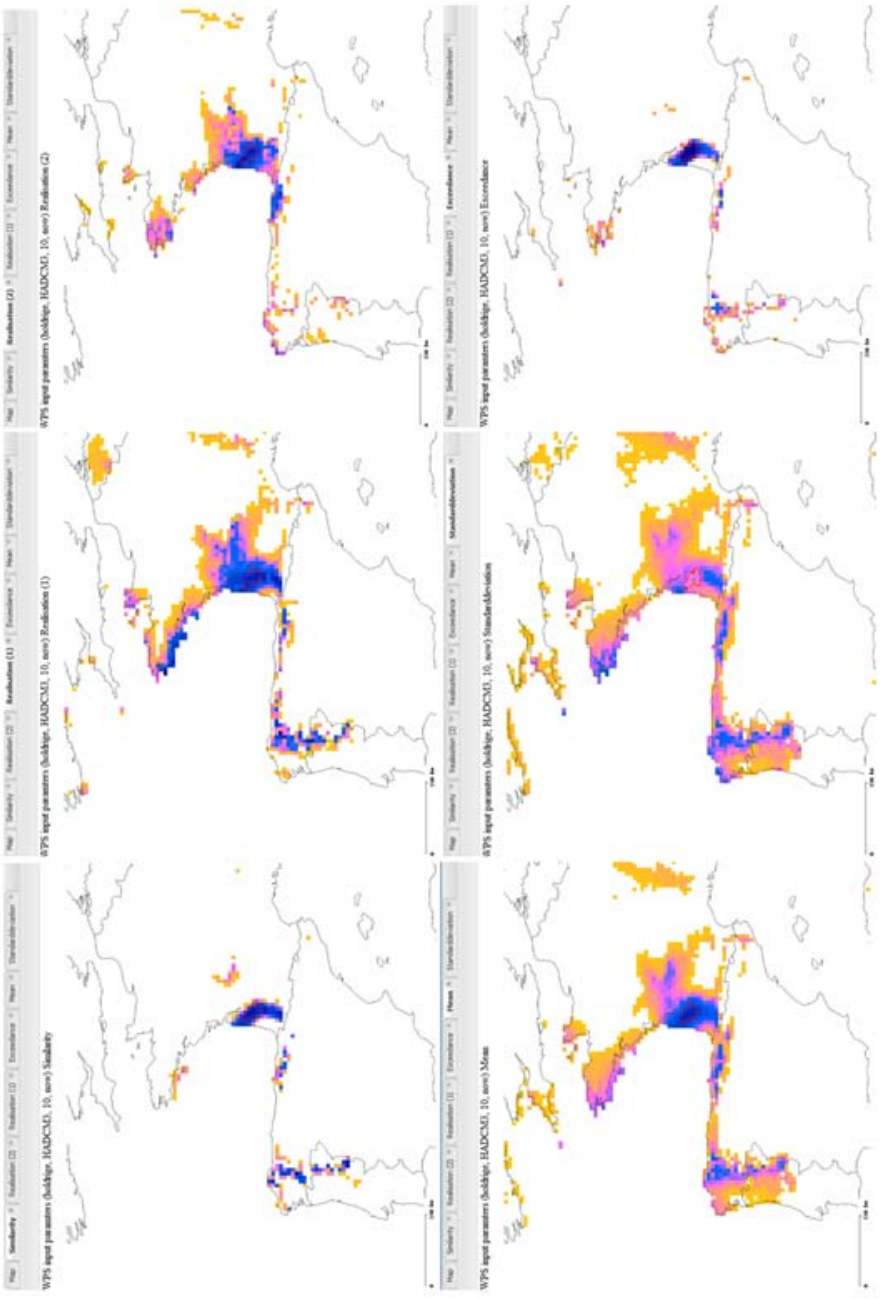

Figure 5 Screen shot of results from eHabitat client - prediction of similarity using the current climatic conditions

Figure 6 shows some results for the forecast climatic conditions, comparing forecast climate for 2050 with the current climatic conditions of the PA Los Alcornocales located in the South of Spain. Showing the same statistics as in Figure 5, we can notice that the conditions in the PA are likely to be rather different than today. The deterministic prediction gives relatively high similarity for some regions on the west coast of Portugal, the north coast of Spain and in France. However, the two realizations shown indicate that the exact location of similar areas will depend highly on the spatial pattern of the uncertainty of the forecast. There are no areas with a high mean forecast similarity, but rather large areas where it is almost equal. This indicates that it would be rather difficult to find the right region for a new PA that could act as a habitat replacement for the species in the PA should they be intolerable to the new climate. It is interesting to note that the mean of the realizations suggests that the future climate of the PA might have some similarity with the current conditions, contrary to the deterministic forecast. The standard deviations of all predictions are relatively high though, and the exceedance probability of a similarity equal to 0.5 is higher than zero just for some small areas scattered around southern Europe and the north coast of Africa. Contrary to the predictions for the current state, we do not see the same reduction in uncertainty around the PA itself, as there is no spatio-temporal correlation between the forecast errors.

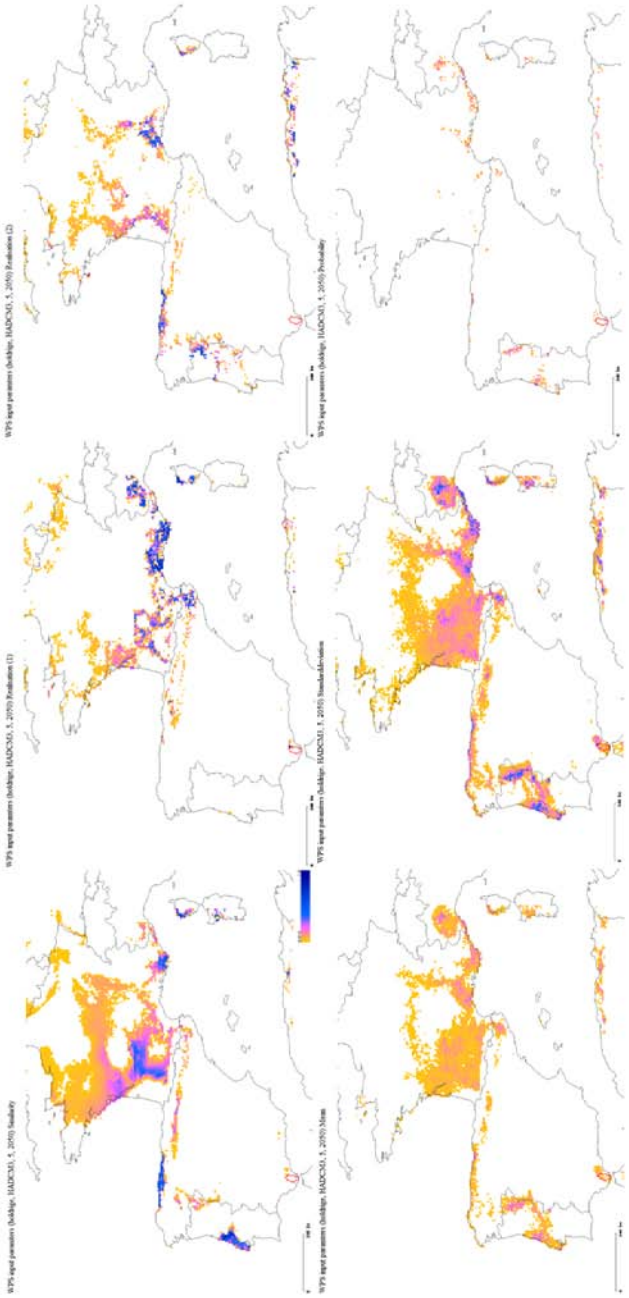

Figure 6 Sreen shot of results from eHabitat client - forecasting scenario

Discussion

We presented some analyses of predictions of similarity using the Mahalanobis distance. Some of the equations presented in the methods section are not used as these are deemed infeasible to use in computation. However, these can still give some help in the quantitative analyses of the results from the more practical analyses using simulations.

The results from the simulations show that it is important to take the uncertainty into account when modeling species distribution. The results usually depend on the application, but we can generally note that high similarities tend to decrease when assuming uncertainty, whereas lower similarities tend to increase slightly. This is due to a smoothing effect of the uncertain approach. For modeling of the current situation, the variability of the result is smaller around the reference region, whereas this cannot be observed for predictions of the forecast similarity.

The example with simulated data shows that it is necessary to consider the number of variables introduced, as this will influence the result. In reality the cross-correlations will be different and not as easy to model as in the example, but it still shows that only variables that truly affect the habitat of species should be included in the modeling.

The method has been implemented in the WPS eHabitat. The Service can be accessed through a web client with a limited data set. The strength of eHabitat is clear when linked with Catalogues and other Web Services. Here, eHabitat is used for ecological forecasting by using climate change data, while it was originally conceived for the identification of similar ecosystems. The possibility to quickly assess different combinations of data sets, and at the same time be informed about the uncertainty of the results is to our knowledge currently not available in any other web service. In this way this service is also another step in the realisation of the Model Web [16].

Acknowledgements

This work is partly funded by the European Commission, under the 7th Framework Programme, by the EuroGEOSS project funded by the DG RTD and by the UncertWEB project funded by the DG INFSO. The views expressed herein are those of the authors and are not necessarily those of the European Commission.

Bibliography

[1] A. Guisan and N.E. Zimmerman, "Predictive habitat distribution models in ecology", Ecological modelling, **135**: p. 147-186,2000.

[2] J.D. Clark, J.E. Dunn and K.G. Smith, "A multivariate model of female black bear habitat use for a geographical information system", Journal of Wildlife Management, **57**: p. 519-526, 1993.

[3] S. T. Knick and D.L. Dyer, "Distribution of black-tailed jackrabbit habitat determined by GIS in southwestern Idaho", Journal of Wildlife Management, **61**(1): p. 75-85, 1997.

[4] J. T. Rotenberry, S.T. Knick and J.E. Dunn, "A minimalist approach to mapping species habitat: Pearson's planes of closest fit", in "Predicting species occurences: issues of accuracy and scale", J.M. Scott, et al., Editors, Island Press: Washington, D. C., USA, 1997.

[5] S. J. Phillips, R.P. Anderson and R.E. Schapire, "Maximum entropy modelling of species geographic distributions", Ecological modelling, **190**: p. 231-259, 2006.

[6] J. T. Rotenberry, K.L. Preston and S.T. Knick, "GIS-based niche modelling for mapping species' habitat". Ecology, **87**: p. 1458-1464, 2006.

[7] C. Calenge, G. Darmon, M. Basille, A. Loison and J.-M. Jullien, "The factorial decomposition of the Mahalanobis distances in habitat selection studies". Ecology, **89**(2): p. 555-566, 2008.

[8] Searle, S., *Linear Models*. 1971, New York: Wiley.

[9] R Development Core Team, R: "A language and environment for statistical computing", R Foundation for Statistical Computing: Vienna, Austria, 2012.

[10] P. H. Hiemstra, E.J. Pebesma, C.J.W. Twenhöfel and G.B.M. Heuvelink, "Real-time automatic interpolation of ambient gamma dose rates from the Dutch Radioactivity Monitoring Network", Computers & Geosciences, **35**(8): p. 1711-1721, 2008.

[11] E. J. Pebesma, "Multivariable geostatistics in S: the gstat package", Computers & Geosciences, **30**: p. 683-691, 2004.

[12] L. R. Holdridge, Determination of world plant formations from simple climatic data". Science, **105**: p. 367-368, 1947.

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

[14] C. W. Thorntwaite, An approach toward a rational classification of climate. Geographical Review, **38**(1): p. 55-94, 1948.

[15] G. Dubois, et al., eHabitat: "a contribution to the model web for habitat assessments and ecological forecasting", in Proceedings of the 34th International Symposium on Remote Sensing of Environment. April 10-15, Sydney, Australia, 2011.

[16] G. Geller and F. Melton, "Looking forward: Applying an ecological model web to assess impacts of climate change", Biodiversity, **9**(3-4): p. 79-83, 2008.