

## THE EFFECT OF DISTANCE CORRECTION FACTOR IN CASE-BASED PREDICTIONS OF VEGETATION CLASSES IN KARULA, ESTONIA

M. Linder\*, L. Jakobson, E. Absalon

Institute of Ecology and Earth Sciences, University of Tartu, Estonia 51014, Vanemuise 46 – [madlili@ut.ee](mailto:madlili@ut.ee)

**KEY WORDS:** Spatial Autocorrelation, Case-based Predictions, Predictive Vegetation Mapping, Machine Learning, Remote Sensing and Map Data, Distance-related Similarity Correction

### ABSTRACT:

The aim of this study was to investigate the applicability of the distance correction parameter (DCP) integrated to the case-based prediction system CONSTUD to reduce the effect of spatial autocorrelation of training data in machine learning process. To achieve this, calculated similarity between observations is decreased by the so-called distance correction value (DCV – the quotient of DCP and distance between two observations). 50 machine learning iterations were carried through in the case of different DCP-s from 0 to 15 000 m using random samples generated from 450 training observations from southern Estonia (Karula National Park and its vicinity). Independent validation samples were used to estimate the effects of the use of each DCP. Machine learning results showed that the Cohen's kappa index of agreement decreased in accordance with the increase of DCP-s. The correspondences of field observations and predicted values followed the same trend. The explanation would be that with the increase of DCP-s successively more observations were rejected as useful ones. Conversely, no considerable decrease in correspondences of the predictions was recognized when DCP was increased. In our case, probably the most useful exemplars were chosen and the less useful ones were left beyond. As a result, scattered and probably spatially and thematically highly representative sample of observations remained. The border might be drawn at DCP from which the number of the in-between distances started to decrease considerably, but the correspondence in validation sample estimations as well as in training sample estimations remained relatively stable.

### 1. INTRODUCTION

This paper is related to the issues of autocorrelation of observations in training samples and to the spatial and thematic representativeness of training data, and also to the overtraining problems in predictive vegetation mapping.

Spatial autocorrelation occurs when locations close to each other have more similar values than those further apart (the values of variables are not independent from each other). Autocorrelation of ecological phenomena may arise for different reasons (see Sokal & Oden, 1978). Positive spatial autocorrelation in moderate distances may accrue from spatial and temporal synchrony of certain abiotic factors that shape particular landscape patterns, e.g., blotched configuration of landscape components. In farther distances, positive autocorrelation may originate from regular variation of environmental gradients and habitat patches. Populations and species may be spatially aggregated due to their dispersal limits caused by different environmental and historical as well as intrinsic organism-specific factors. Among other reasons causing spatial autocorrelation in (predictive) models, is omitting an important variable from the model, observation biases (variance in data collection, sampling and mapping) (Dormann et al., 2007), etc.

Spatial autocorrelation may be interpreted as intrinsic feature of a phenomenon providing additional information for spatial analysis. When spatial autocorrelation occurs, the values of variables are predictable on the basis of the values of the same variable in other locations. Luoto et al. (2005) found that performance of species-climate models depends on geographical attributes of the species, including spatial autocorrelation. They also found that butterfly species with

more aggregated occurrence pattern (expressing high spatial autocorrelation) were better predicted compared to the species with scattered distribution (exhibiting low autocorrelation).

However, the presence of spatial autocorrelation is frequently a disadvantage for hypothesis testing and prediction, because it violates one of the main assumptions of standard statistical analyses that residuals are independent and identically distributed (Dormann et al., 2007). The presence of positive spatial autocorrelation in model residuals (spatial dependency) may bias parameter estimates and can increase the likelihood of type I statistical error (Betts et al., 2006).

Since spatially close locations/observations tend to be similar due to spatial autocorrelation, they predict each-other with great accuracy. As a result, deceptively high prediction accuracy (overtraining) occurs and the application of this set of training observations for more distant locations would not be so reliable.

Accounting for spatial autocorrelation should increase prediction versatility. Taking into account autocorrelation is crucial when image data are classified, because estimations of classification accuracy that compare prediction and actual situation pixel by pixel, tend to overestimate results due to the autocorrelation of the pixel values (Muchoney & Strahler, 2002).

A variety of widespread statistical tools have been developed to correct for the effects of spatial autocorrelation in species distribution data. Dormann et al. (2007) presented different statistical approaches that efficiently accounted for spatial autocorrelation in analyses of spatial data. Most of the spatial modeling techniques they tested on spatially autocorrelated simulated data showed good type I error control and precise

---

\* Corresponding author.

parameter estimates. Accounting for autocorrelation via autologistic models has become common (e.g., Augustin et al., 1996; Osborne et al., 2001; Luoto et al., 2005). It has been shown that including spatial autocovariates improves model prediction success (Augustin et al. 1996; Osborne et al., 2001; Knapp et al., 2003; Betts et al., 2006). Generalized estimating equations (GEE – an extension of generalized linear models) have been used by Augustin et al., 2005, Carl & Kühn 2007, etc. The use of GEE models reduced the autocorrelation of the residuals considerably indicating effectively removed spatial dependency. Though, Diniz-Filho et al. (2003) concluded that ignoring spatial autocorrelation does not cause problems necessarily in all analyses.

The aim of this study was to introduce and test the usability of the distance correction parameter integrated to the case-based prediction system CONSTUD for extenuating the effect of the autocorrelation in predictions of spatial phenomena (in this case – vegetation classes).

## 2. METHODOLOGY

### 2.1 Study Area and Field Data

450 training observations from southern Estonia (Karula National Park and its vicinity; Figure 1) were gathered mostly during the summers 2007 and 2008, partly during the inventories from 2001 to 2007. EUNIS classes (Davies et al., 2004) were used as a predictable variable.

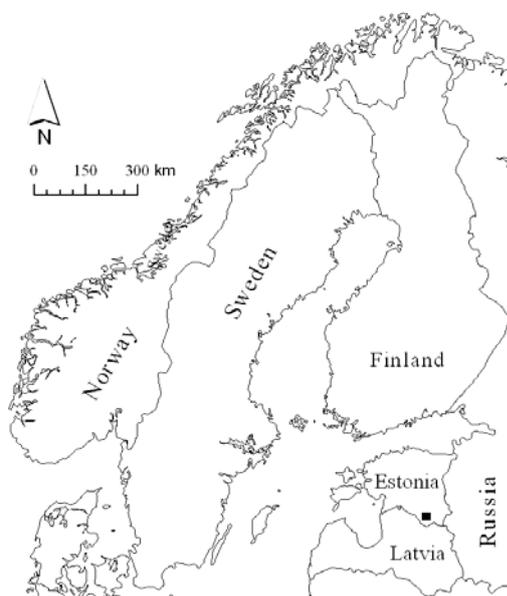


Figure 1. Location of the study area (black square).

### 2.2 Remote Sensing and Map Data

Data layers of explanatory variables were: rasterized 1:10 000 Estonian base map and 1:10 000 digital soil map, Landsat 5 TM satellite images (scenes 186-19 and 186-20) of 21st of May 2007 and 9th of August 2007, and the orthophotos from the year 2005. Layers for red, green, blue, yellow, hue, saturation and lightness were derived from orthophotos. In the case of satellite images, the NDVI layers were derived from the near-

infrared and red channel values. In addition, the Baltic SRTM30 (*Shuttle Radar Topography Mission*) elevation model was used. The data layers were prepared according to the prerequisites for the application of CONSTUD (CONSTUD, 2009) using ArcGIS 9, Idrisi Andes, LSTATS (LSATS, 2009) and an original application for rasterizing the soil map.

### 2.3 Case-based Prediction System CONSTUD

The case-based (Aha, 1998; Remm, 2004) machine learning and prediction system created in the University of Tartu by Kalle Remm was used (more details and case studies in Linder et al., 2008; Remm & Remm, 2008; Remm et al., 2009; Remm & Remm, 2009; Tamm & Remm, 2009; CONSTUD, 2009). CONSTUD was used for: 1) calculating the pattern indices in training locations from map and image data (explanatory variables), 2) machine learning – iterative search for the best set of feature weights of the observations and the best observations (exemplars), 3) predictions of vegetation classes.

Decisions are made on the basis of similarity between studied cases and predictable sites in CONSTUD. Similarity between observations is calculated as a weighted average of partial similarities of single features (further details: Linder et al., 2008; Remm, 2004). During machine learning process, goodness-of-fit of predictions is estimated using leave-one-out cross validation (the predicted value for every observation is calculated using all exemplars except this particular one), and in the case of multinomial variable (like vegetation classes), Cohen's kappa index of agreement is used to measure the correspondence of predictions to observations.

### 2.4 Distance-related Similarity Correction

Distance correction parameter (DCP) is integrated to the system CONSTUD to reduce the effect of spatial autocorrelation in training data in machine learning process. DCP regulates the extent of reciprocal prediction of close observations by decreasing the calculated similarity between observations in proportion to the inverse distance between them. The extent of decrease is regulated by DCP (in meters) chosen by the user. Distance correction value (DCV) is calculated as the ratio of DCP and distance between two observations. Then calculated similarity between these values is corrected. Corrected similarity value (CSV) is gained as calculated similarity (from 0 to 1) minus DCV. If the distance between two observations is equal to or less than DCP, then CSV is set to 0 even if the calculated similarity is 0.9. The closer are observations, the higher is the rate the similarity between them is corrected. In this study, DCP-s from 0 to 20 000 m were tested.

### 2.5 Calculations

First, explanatory variables (spatial pattern indices) from image and map data layers were calculated. Then, 50 machine learning iterations were carried through in the case of each selected DCP in two stratified random samples generated from 450 training observations. One of the samples was first used as a training sample and the other as a validation sample. Then the roles were exchanged and the results were averaged. Independent validation samples were used to give the estimation for the use of each particular DCP. Finally, the correspondences for predictions (the proportion of coincident observations among all observations) were calculated.

### 3. RESULTS

The results showed that the Cohen's kappa index of agreement continually decreased with the increase of the DCP-s (Figure 2).

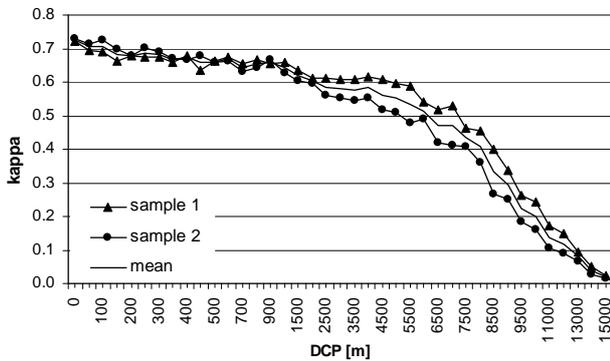


Figure 2. Kappa values gained during machine learning iterations of two random samples using different distance correction parameters (DCP-s).

The correspondences of field observations and predicted values in machine learning followed the same trend (Figure 3). The reason is probably the fact that with the increase of DCP lower weights were attributed to successively more observations.

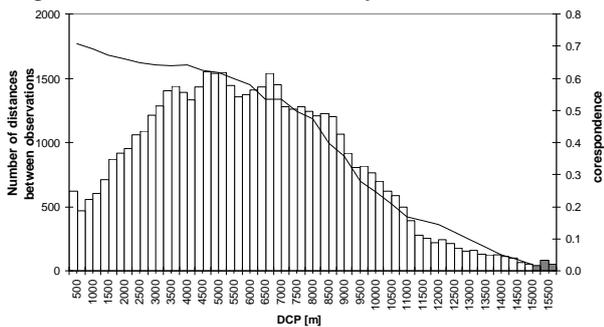


Figure 3. Line – correspondence of field observations and classes estimated during machine learning iterations (mean of two samples). White columns – distribution of all the distances between all observations (total of 50 400). Grey columns – distances between observations used in the last samples (in the case of the highest DCP – 15 000 m – 179 distances, i.e., 0.36% of all distances were comprised).

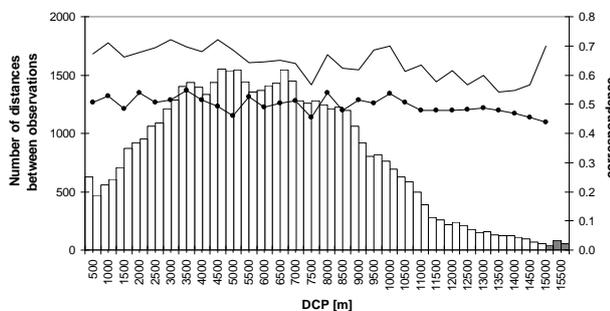


Figure 4. Line with dots – correspondences when estimating validation samples (mean of the two samples). Regular line – correspondences when estimating machine learning samples (mean of the two

samples). White and grey columns – see Figure 3 caption.

Conversely, in the case of predictions, no considerable decrease in correspondences was recognized when DCP was increased (Figure 4). Furthermore, the lines of correspondences approached to each other (Figure 4, 5).

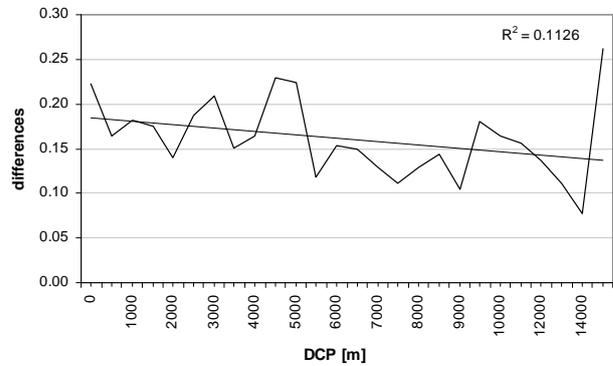


Figure 5. Differences between correspondences of estimated samples that were fitted during machine learning iterations and those of validation samples.

Samples from within the range of spatial autocorrelation may be inefficiently large (e.g., unduly time-consuming), because observations with spatially autocorrelated values will probably add little independent information (Dormann et al., 2007). It might be suggested that into the samples of our study, the most useful exemplars were chosen and the less useful ones were left beyond by CONSTUD. As a result, dispersed and probably highly (spatially and thematically) representative compact samples of observations remained (Figure 6). The border might be drawn at DCP from which the number of the in-between distances started to decrease considerably, but the correspondence in validation sample estimations as well as in training sample estimations remained relatively stable – in this case, at DCP of somewhere between 8500 and 9000 m (Figure 4).

### 4. CONCLUSIONS

Relying upon the results of this study, the use of distance correction parameter in case-based prediction and machine learning system CONSTUD gives a presumably thematically and spatially representative training sample which in turn reduces or removes the effect of autocorrelation and noise in data. This enables reducing the time expended on calculations of predictions. However, as far as only 450 observations were used (furthermore, these were divided into training and validation samples), wider interpretations could be biased. Using higher amount of field observations might give converse results, or might just increase the time for calculations. Also, expanding the area from where field data are gathered from could have unpredictable effects, due to the unique character of different landscape regions or for any other reason.

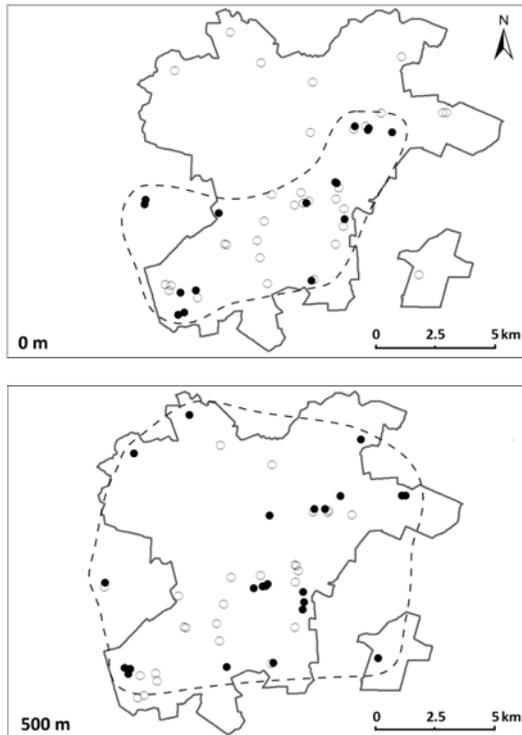


Figure 6. The effect of the use of distance correction parameter (DCP) in the case of coniferous forest observations in and around Karula National Park (within grey boundary line). Black dots – used exemplars, transparent dots – cases that turned out to be not very useful during machine learning iterations. Upper figure – DCP = 0 m, bottom figure – DCP = 500 m. Compared to the case when DCP was not used, the used exemplars are more dispersed (area within dashed line) (Linder et al., 2009, modified).

## REFERENCES

- Aha, D. W., 1998. The omnipresence of case-based reasoning in science and application. *Knowledge-Based Systems*, 11, pp. 261–273.
- Augustin N.H., Kublin E., Metzler B., Meierjohann E., von Wuhlisch G., 2005. Analyzing the spread of beech canker. *Forest Science*, 51, pp. 438–448.
- Augustin N.H., Mugglestone M.A., Buckland S.T., 1996. An autologistic model for the spatial distribution of wildlife. *Journal of Applied Ecology*, 33, pp. 339–347.
- Betts, M. G., Diamond, A.W., Forbes, G.J., Villard, M.-A., Gunn, J.S., 2006. The importance of spatial autocorrelation, extent and resolution in predicting forest bird occurrence. *Ecological Modelling*, 191, pp. 197–224.
- Carl, G., Kühn, I., 2007. Analyzing spatial autocorrelation in species distributions using Gaussian and logit models. *Ecological Modelling*, 207, pp. 159–170.
- CONSTUD, 2009. <http://www.geo.ut.ee/CONSTUD> (accessed 13 Dec. 2009).
- Davies, C. E., Moss, D., Hill, M. O., 2004. EUNIS Habitat Classification. Revised 2004. [http://eunis.eea.europa.eu/upload/EUNIS\\_2004\\_report.pdf](http://eunis.eea.europa.eu/upload/EUNIS_2004_report.pdf) (accessed 15 Oct. 2009).
- Diniz-Filho J.A.F., Bini L.M., Hawkins B.A., 2003. Spatial autocorrelation and red herrings in geographical ecology. *Global Ecology & Biogeography*, 12, pp. 53–64.
- Dormann, C. F., McPherson, J. M., Araújo, M. B., Bivand, R., Bolliger, J., Carl, G., Davies, R. G., Hirzel, A., Jetz, W., Kissling, W. D., Kühn, I., Ohlemüller, R., Peres-Neto, P. R., Reineking, B., Schröder, B., Schurr, F. M., Wilson, R., 2007. Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography* 30(5), pp. 609–628.
- Knapp, R.A., Matthews, K.R., Preisler, H.K., Jellison, R., 2003. Developing probabilistic models to predict amphibian site occupancy in a patchy landscape. *Ecological Applications*, 13, pp. 1069–1082.
- Linder, M., Remm, K., Absalon, E., 2009. The utility of the machine learning and prediction system CONSTUD. In: Mander, Ü., Uuemaa, E., Pae, T. (Eds.). *Uurimusi eestikeelse geograafia 90. aastapäeval. Publicaciones Instituti Geographici Universitatis Tartuensis*, 108, pp. 52–62. Tartu: Tartu University Press.
- Linder, M., Remm, K., Proosa, H., 2008. The application of the concept of indicative neighbourhood on Landsat ETM+ images and orthophotos, using circular and annulus kernels. In: Ruas, A., Gold, C. (Eds.). *Proceedings of the 13th International Symposium on Spatial Data Handling*, Montpellier, France, 23rd–25th June, pp. 147–162. Springer.
- LSTATS, 2009. <http://www.geo.ut.ee/LSTATS> (accessed 13 Dec. 2009).
- Luoto M., Poyry J., Heikkinen R.K., Saarinen K., 2005. Uncertainty of bioclimate envelope models based on the geographical distribution of species. *Global Ecology and Biogeography*, 14, pp. 575–584.
- Muchoney, D. M., Strahler, A. H., 2002. Pixel- and site-based calibration and validation methods for evaluating supervised classification of remotely sensed data. *Remote Sensing of Environment*, 81, pp. 290–299.
- Osborne P.E., Alonso J.C., Bryant R.G., 2001. Modelling landscape-scale habitat use using GIS and remote sensing: a case study with great bustards. *Journal of Applied Ecology*, 38, pp. 458–471.
- Remm, K., 2004. Case-based predictions for species and habitat mapping. *Ecological Modelling*, 177(3–4), pp. 259–281.
- Remm, K., Linder, M., Remm, L., 2009. Relative density of finds for assessing similarity-based maps of orchid occurrence. *Ecological Modelling*, 220(3), pp. 294–309.
- Remm, K., Remm, L., 2009. Similarity-based large-scale distribution mapping of orchids. *Biodiversity and Conservation*, 18(6), pp. 1629–1647.
- Remm, M., Remm, K., 2008. Case-based estimation of the risk of enterobiasis. *Artificial Intelligence in Medicine*, 43(3), pp. 167–177.

Sokal, R. R., Oden, N., 1978. Spatial autocorrelation in biology. 1. Methodology. *Biological Journal of the Linnean Society*, 10, pp. 199–228.

Tamm, T., Remm, K., 2009. Estimating the parameters of forest inventory using machine learning and the reduction of remote sensing features. *International Journal of Applied Earth Observation and Geoinformation*, 11(4), pp. 290–297.

#### **ACKNOWLEDGEMENTS**

The authors are especially grateful to the creator of the system CONSTUD – Kalle Remm – for his helpful comments. We also thank all the people whose field observations were used in this study and Olivia Till from Karula National Park who provided us with additional observations from inventories. The maps and orthophotos were used according to the Estonian Land Board licenses (107, 995, 1350). The research was supported by the Estonian Ministry of Education and Research (SF0180052s07, SF0180049s09) and by the University of Tartu (BF07917).