20th International Congress on Modelling and Simulation, Adelaide, Australia, 1–6 December 2013
www.mssanz.org.au/modsim2013

A Naive Bayes classifier for modeling distributions of the common reed in Southern Finland

A. Altartouri[a] and A. Jolma[a]

[a] *Department of Civil and Environmental Engineering, School of Engineering, Aalto University, Finland*
Email: anas.altartouri@aalto.fi

Abstract: The field of species distribution and habitat suitability modeling has witnessed significant advancements in a number of aspects. One area that received much attention is the statistical underlying models in these studies. As data becoming bulky and prediction is often the goal of modeling, Data Mining and Machine Learning methods are becoming favorable in providing the underlying probability models for species distribution studies. Machine Learning encompasses a wide range of classification techniques, among others, with various capabilities. Although a number of techniques were presented and applied in species distribution modeling, many remain still untested. We here examine the potential of the Naive Bayes classification method, a widely and successfully applied technique in a number of fields, for modeling the common reed *Phragmites australis* distributions. We developed a Naive Bayes classifier to predict occurrences of *Phragmites australis* in a site on the Southern Finnish coast. We also tested the potential of the classifier to provide input to a cellular automaton for modeling the spread of *Phragmites australis*. The results suggests that the Naive Bayes classifier has significant potential in predicting species occurrences and providing transition rules for the dynamic modeling of species distributions.

Keywords: *Species distribution models, Phragmites australis, Machine Learning, Naive Bayes, cellular automata, Gulf of Finland*

1. INTRODUCTION

Species distribution modeling (SDM) is an active area of research that over years has witnessed significant advancements in a number of aspects. In their editorial, Zimmermann and others (2010) highlighted new trends in SDM. Much methodological advancement in the last decade was achieved in the underlying statistical bases of SDM. Advancement, though to a lesser extent, occurred also in different directions such as the use of dynamic models. Machine Learning (ML) methods were introduced to the field and found to improve prediction (Elith *et al.*, 2006). The growing interest in these methods can be attributed, in addition to the improved prediction, to their ability to model complex relationships in ecological data without the need to meet a number of restricting assumptions often found in the parametric approaches (Hochachka, *et al.*, 2007; Olden *et al.*, 2008).

There is a wide range of methods available for classification in the fields of Machine Learning and Data Mining. These methods vary in their strengths and weaknesses, and perform differently in different fields and with different data sets. A number of ML methods for ecological modeling were compared in the literature (*e.g.* Elith *et al.*, 2006; Olden *et al.*, 2008). One of the classification methods that is widely applied in various fields is the Naive Bayes (NB) classification (Yang and Webb, 2009, Larsen, 2005). It is a simple and effective classifier that is easy to build and understand, and it outperforms a number of other types of classifiers in many tasks (Yang and Webb, 2009; Larsen, 2005). However, the potential of NB classification in species distribution modeling remains untested. Moreover, there is a need for novel solutions combining robust prediction methods and dynamic modeling in the study of species distributions.

In this paper, we address these two points, namely the use of ML methods combined with a Cellular Automaton for dynamic modeling of species distributions. More specifically, the aim of this paper is to examine the potential of NB classification in modeling distributions of the common reed in a site in Southern Finland. We test the potential of NB classifier to assess the suitability of geographic locations for the reed establishment. We also examine the ability of NB classifier to provide input to a cellular automaton for dynamic modeling and prediction of future distributions of the common reed.

The species in question is the common reed, *Phragmites australis*. It is a native species that has markedly spread into many places along the Finnish coasts in the last decades. It is influencing the distribution and abundance of a number of other species and lowering the value of the impacted coastal and archipelago properties (Ikonen and Hagelberg, 2007). A model capable of defining areas susceptible for *Phragmites* takeover and predicting its future distribution is therefore needed for the planning and management of the area. Human disturbance and urban development in coastal areas and nitrogen pollution are found to facilitate colonization of clear shorelines by *Phragmites*. (King *et al.*, 2007). Once established, *Phragmites* proliferates mostly vegetatively by rhizomes and forms large colonies (Bart and Hartman, 2003).

2. MATERIALS AND METHODS

2.1. Study area and data

The methods presented in this paper were applied on the case of *Phragmites* expansion in Svartbäck (Purola), a small site on the Southern Finnish coast of the Gulf of Finland (GOF). The site is about 30 km^2, near to the outlet of River Kymijoki which is one of the major rivers flowing into the GOF at Ruotsinpyhtää (Figure 1). Available datasets included (a) grids of *Phragmites* distribution in the site in 2003 and 2006, delineated from aerial photographs by the Finnish Environment Institute (SYKE) and rasterized as binary grids of 1/0 denoting the reed presence/absence; (b) a bathymetry grid (in decimeters) derived from contour lines and depth measurement points; (c) a grid of relative water openness given by the abstraction of fetch lines in 36 directions; and (d) a grid of the Euclidean distance to the closest river mouth (in meters).

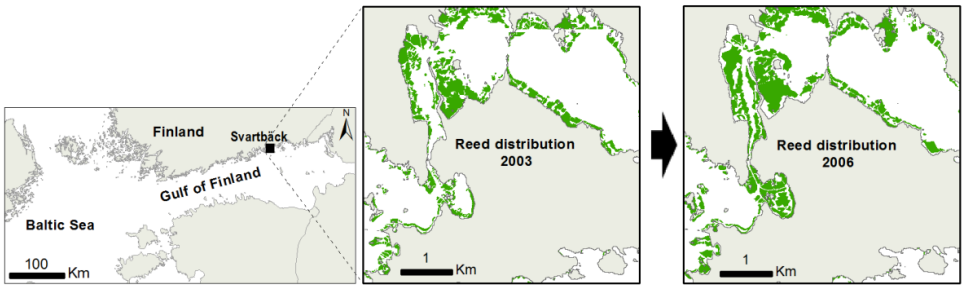

Figure 1. Location of study area, and distribution of the common reed in 2003 and 2006.

2.2. Exploratory analyses

Analyses were performed to examine the relationships between explanatory variables and (a) the presence or absence of *Phragmites* and (b) the expansion or disappearance of *Phragmites* from certain locations. The first is concerned with the occurrence of the species in the geographic space while the second examines the dynamics of *Phragmites* colonies in space and time. The indicated causes of the invasive expansion of *Phragmites* and the mechanism by which it expands mentioned in the Introduction was the starting point for variable selection. Due to the unavailability of sufficient data of nutrient content in the sediment, the proximity to a river mouth (denoted '*dis*') was used as a surrogate variable. Two variables are examined in addition; the depth of water (denoted '*dep*') and the relative openness of water (denoted '*opn*'). The depth of water is observed to influence the offshore-extent to which *Phragmites* colonies advance. The complexity of the Finnish coastline requires the incorporation of the shore openness variable. The formation of the shoreline together with the presence of thousands of islands is key to the coastal processes in such areas (Ekebom *et al.*, 2003). The openness of a location is given by the normalized sum of fetch lines originated from that location to the first interrupting object (an island) in 36 directions. Finally, in order to account for the neighborhood effect, the count of reed-occupied cells in a Moore neighborhood window was analyzed.

All analyses were performed using a raster data model. Histograms were produced to provide insights on the presence or absence of *Phragmites* at different values of the explanatory variables. Analysis of the expansion or disappearance of *Phragmites* considers the state transition of cells between two successive time steps (*t* and *t+1*) which can take one of four possible cases; cells that were free of *Phragmites* in both time steps (denoted 0->0), cells to which *Phragmites* has expanded by *t+1* (denoted 0->1), cells from which *Phragmites* has disappeared by *t+1* (denoted 1->0), and cells that were occupied by *Phragmites* in both time steps (denoted 1->1). A grid holding the number of reed-occupied cells within a Moore neighborhood window at *t* was cross-tabulated with a grid holding the state transition, resulting in a two-way table. Entries in this table give the proportion of each case of state transition for each possible number of reed-occupied cells at the initial time step *t*. The table was then divided based on the initial state; the '0->0' and '0->1' cases formed a table that reflects on the expansion of *Phragmites*, and the '1->0' and '1->1' cases formed a table that reflects on the disappearance of *Phragmites*. Stacked bar charts were then produced to illustrate these relationships and depict the effect of the neighborhood composition on the state of cells in the following time steps.

2.3. Naive Bayes classification

A Naive Bayes classifier is a simple probabilistic classifier based on Bayes' Theorem. It is a supervised algorithm for the prediction of a categorical (discrete) response variable. Bayesian classifiers attempt to predict a discrete class *c* from a finite number of classes *C* given values a_1 through a_n, of attributes A_1 through A_n. This can be written using Bayes' rule as:

$$Pr(C = c|A_1 = a_1, A_2 = a_2, \cdots, A_n = a_n) = \frac{Pr(A_1=a_1, A_2=a_2, \cdots, A_n=a_n|C=c)\ \Pr(C=c)}{Pr(A_1=a_1, A_2=a_2, \cdots, A_n=a_n)} \quad (1)$$

The prior probability *Pr(C=c)* can be easily estimated from the labeled training data. The denominator of the above equation is not dependent on *C* and is therefore irrelevant for classification. The difficulty stems from the infeasible number of parameters to be estimated when expanding the class conditional probability in the numerator as follows (value annotations are omitted):

$$Pr(A_1, A_2, \cdots, A_n|C) = Pr(A_1|A_2, \cdots, A_n, C)\ Pr(A_2|A_3, \cdots, A_n, C)\ Pr(A_3|A_4, \cdots, A_n, C)\ \cdots\ Pr(A_n|C) \quad (2)$$

NB simplifies the above equation by assuming that attributes are conditionally independent, that is, the outcome of an attribute is independent of the outcome of all other attributes given the class value. Equation (2) is therefore simplified as:

$$Pr(A_1, A_2, \cdots, A_n|C) = Pr(A_1|C)\ Pr(A_2|C)\ Pr(A_3|C)\ \cdots\ Pr(A_n|C) \quad (3)$$

Each of the factors in Equation (3) can be estimated from the labeled data as:

$$Pr(A_i = a_i|C = c) = \frac{Count(A_i=a_i\ \wedge\ C=c)}{Count(C=c)} \quad (4)$$

The classifier assigns an unlabeled instance to the class with the highest probability estimate. Despite the obvious violation of the independence assumption in most cases, the NB classifier is reported to behave successfully in a wide range of classification tasks (Yang and Webb, 2009; Cheng and Greiner, 1999). For a more comprehensive description of the NB algorithm see *e.g.* Mitchell (2005).

We trained NB classifiers using a sample of 10% of the data. The sample was drawn using a conditional Latin Hypercube Sampling procedure (Minasny and McBratney, 2006). With this procedure, the sample was

marginally maximally stratified for each variable, including the class variable, the explanatory variables, and the geographic space given by x and y coordinates. This is essential for getting as accurate prior and conditional probabilities as possible for the NB classifier. The trained classifier was then used to classify the unlabeled instances (representing grid cells) in the study area. The probability of each cell to be classified as reed-occupied was outputted as an estimate for the suitability of that cell for hosting *Phragmites*.

2.4. Cellular automata modeling

A cellular automaton (CA) is a discrete model where space is represented by a grid of cells with a finite number of states (Fonstad, 2006). A CA is composed of five elements; *space* is represented as a grid, which is usually a regular grid of square cells; each cell in the grid has one of predefined *discrete states*; during the simulation all cells evolve in discrete *time steps*; the evolution of cells is determined by a set of *transition rules* which in turn are dependent on the specified *neighborhood configuration* (Ménard and Marceau, 2005).

We developed a CA to run a dynamic spread model of *Phragmites* in the study site. The model was initiated with the distribution of *Phragmites* in 2003 (C_{2003}) and run for three generations, where each generation represents a year, to predict the distribution in 2006 (C_{2006}). The neighborhood window was set to 3x3, and the model cell size to 2 m. The CA transition probabilities were given by a NB classifier. The classifier was trained to predict the state (class) of a cell in the next time step (C_{t+1}) given its state at the current time step (C_t), the composition of its neighborhood states at the current time step (N_t) (given by the count of reed-occupied cells), and the explanatory variables (depth of water, openness, and distance to a river mouth). The CA model is based on a previously developed model by the authors (Altartouri and Jolma, 2012), modified by incorporating suitability factors in the prediction of future distributions of *Phragmites* colonies, and simplified by accounting for the composition of the neighborhood state rather the state of individual cells within the neighborhood window.

2.5. Evaluation

The performance of the classifiers was evaluated using 10-fold cross-validation. The suitability map was compared on a cell-by-cell basis with the observed *Phragmites* distribution in 2003 and 2006. The predicted distribution of 2006 from different models was compared similarly with the actual distribution of year 2006. Accuracies from both comparisons were given by the proportion of cells correctly classified. With a boolean-valued class variable, the fraction of cells correctly classified as reed-occupied (*i.e. Pr(Class = 1) >= 50%*) and the fraction of cells correctly classified as reed-free (*i.e. Pr(Class = 1) < 50%*) are, respectively, referred to as the true positive (sensitivity) and the true negative (specificity) rates. The other two proportions are the error rates referring to cells assigned to one of the classes while observed otherwise.

3. RESULTS AND DISCUSSION

3.1. Exploratory analyses

The correlations between the occurrence of *Phragmites* in 2003 and the depth of water, openness of water, and proximity to river mouths were -0.328, -0.236, and -0.243, respectively. They were slightly higher in 2006 with coefficients of -0.353, -0.262, and -0.272, respectively. The correlations are all negative, as one would expect (depth is given in positive values). For a clearer picture, Figure 2 illustrates the distribution of each variable for reed-occupied (in 2003 or 2006) and reed-free cells. No relationships were found between the state transition and the explanatory variables (no graphs are presented).

These results are in line with the hypothesis that shallow and close water areas nearby river mouths are more susceptible for the establishment of *Phragmites*. The comparable correlation of the occurrence of *Phragmites* with the variables in both years suggests that *Phragmites* colonies advanced into areas with similar location characteristics with respect to the investigated variables. The discrimination between variable distributions of reed-occupied and reed-free locations, although not very pronounced, indicates the potential of the selected variables for the prediction of *Phragmites* occurrence (as shown by the classification results below).

The neighborhood of a location is found to be influencing its state in the following time steps. Figure 3(a) shows the proportion of cells to which *Phragmites* expanded by 2006 after being clear of *Phragmites* in 2003 to those cells that were free of *Phragmites* in both time steps (case 0->1 to 0->0), for each case of neighborhood composition (given by the count of reed cells in the neighborhood). On the other hand, Figure 3(b) illustrates the proportion of reed-occupied cells in 2003 that became clear of *Phragmites* in 2006 to those that were occupied by *Phragmites* in both time steps (case 1->0 to 1->1), for each case of neighborhood composition. The graphs suggest that clear water areas are susceptible for *Phragmites* colonization if their

surroundings are already taken by *Phragmites*. On the contrary, *Phragmites* patches seem to disappear from locations where not enough adjacent *Phragmites* beds exist.

The influence of the neighborhood on the expansion and disappearance of *Phragmites* from a location is likely due to the mechanism by which it disperses. New shoots of reed are expanding by rhizomes to abutting sites. However, this neighborhood effect incorporates as well an exogenous component from the location's suitability since the explanatory variables are spatially autocorrelated, which means that presence of *Phragmites* in a location indicates the suitability of its neighborhood for hosting *Phragmites*.

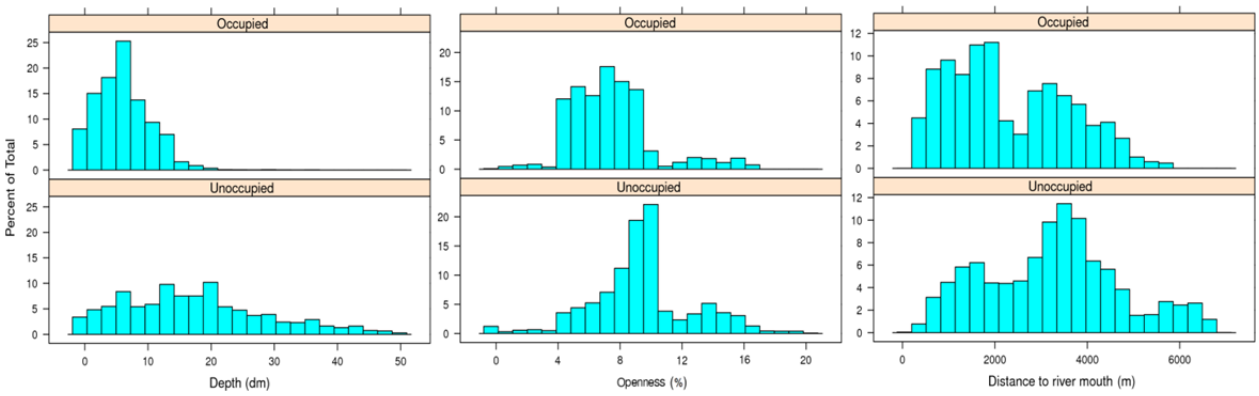

Figure 2. Distributions of suitability variables for reed-occupied cells (in 2003 or 2006) and reed-free cells.

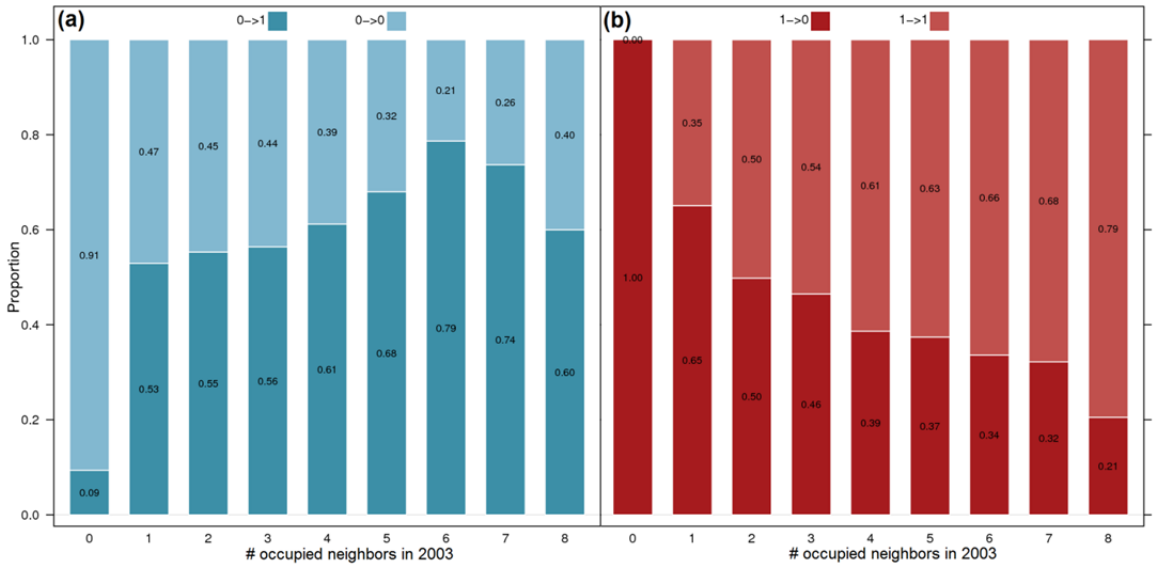

Figure 3. Proportion of each case of state transition between 2003 and 2006 for each neighbourhood composition.

3.2. Classification

The NB classifier incorporating three explanatory variables was used to classify cells in the study area with respect to the presence or absence of *Phragmites*. The probability assigned for cells to be classified as reed-occupied are used to produce the map in Figure 4(a). The map provides a suitability assessment of each cell for *Phragmites* establishment. The accuracy of the classifier was 0.753. The suitability map is compared with *Phragmites* distributions of 2003 and 2006. Cells with *Pr(Class = 1) >= 50%* were considered as *Phragmites* cells. Table 1 lists the sensitivity and specificity resulted from this comparison.

The classifier performed relatively well considering the number of variables used for classification. Unfolding this accuracy (0.753) by considering the classifier's sensitivity and specificity in Table 1 shows that its ability to predict the presence of *Phragmites* is lower than predicting its absence. This can be attributed to the fact that the number of cases (cells) where *Phragmites* is absent in the study area is almost three times more than the number of cases where it is present, causing the classifier to assign a greater prior (background) probability for the absence case.

Table 1. Sensitivity and specificity of classifiers used in the suitability assessment and the CA simulation.

	Assessment of location suitability $Pr(C \mid dep, opn, dis)$		CA simulation of 2006 distribution $Pr(C_{t+1} \mid dep, opn, dis, C_t, N_t)$
Compared with	Observed 2003 distribution	Observed 2006 distribution	Observed 2006 distribution
Sensitivity	*0.639*	*0.623*	*0.696*
specificity	*0.787*	*0.816*	*0.915*

3.3. Dynamic modeling

Table 2 lists the accuracy of classifiers used to predict *Phragmites* distribution of 2006 (C_{2006}). These classifiers use different input to perform the classification task; the first uses suitability variables without incorporating time- or space-dependent attributes; the second adds to the suitability variables the state (or class) of the instance in 2003 (C_{2003}); and the third adds to the latter the neighborhood state in 2003 (N_{2003}). The predicted *Phragmites* distribution of 2006 is presented in Figure 4(b). The sensitivity and specificity of the CA simulation of *Phragmites* distribution of 2006 compared with truth are shown in Table 1.

The accuracy of classification was enhanced by accounting for time autocorrelation. Classifiers predicted the states of cells in 2006 (C_{2006}) more accurately by considering their states in 2003 (C_{2003}) together with the suitability variables. Further performance enhancement was achieved by the incorporation of the spatial neighborhood composition (N_{2003}) in the classification. These results are in line with the literature emphasizing the importance of accounting for autocorrelation in modeling species distributions (*e.g.* Dormann *et al.*, 2007). Although nonparametric, accounting for temporal and spatial autocorrelation through the incorporation of the state of cells and their neighborhood compositions in the previous time step in the CA resulted in a more accurate prediction of future distributions of *Phragmites* colonies.

Table 2. Accuracy of classifiers trained with different input.

Classifier	$Pr(C_{2006} \mid dep, opn, dis)$	$Pr(C_{2006} \mid dep, opn, dis, C_{2003})$	$Pr(C_{2006} \mid dep, opn, dis, C_{2003}, N_{2003})$
Accuracy	*0.765*	*0.813*	*0.844*

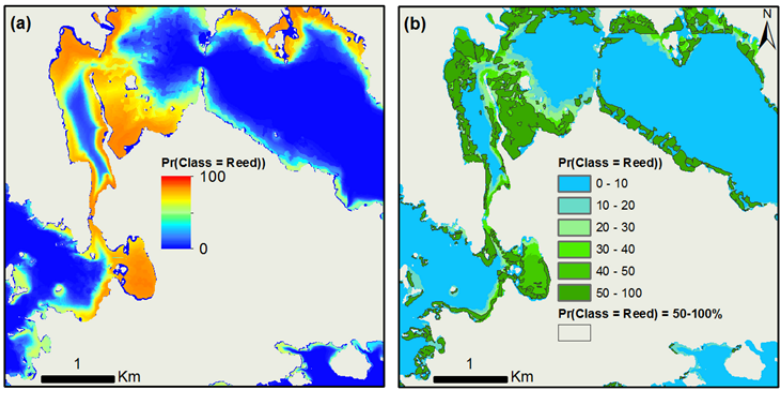

Figure 4. Suitability map (left) and predicted *Phragmites* distribution of 2006 (right).

4. CONCLUSIONS AND FUTURE WORK

Our results concur with earlier work which has shown that ML methods are useful in modeling species distributions due to their ability to improve prediction even with surrogate variables which are easy to obtain. In many cases, the goal of species modeling exercise is to predict species occurrences rather than to explain causal drivers of their distributions. We tested a classification method which has earlier been used successfully in a number of cases, namely the Naive Bayes classification, for modeling *Phragmites* distribution. The method exhibited good performance and high potential for such modeling tasks. The NB classifier produced suitability maps with an acceptable accuracy. It was also useful in learning transition rules for the CA that was used in the dynamic modeling of *Phragmites* distribution. The NB classifier is available in a number of free and open source software packages, *e.g.* R, Weka and Orange[1], and is moderate in

[1] R (www.r-project.org), Weka (www.cs.waikato.ac.nz/ml/weka), and Orange (orange.biolab.si).

computation and memory requirements. An obvious limitation of the NB classifier is its inability to model interactions between explanatory variables which is, however, required in a number of cases.

A number of directions for future work arise. Comparing NB classification with other methods, such as boosted regression trees (Elith *et al*., 2006) would be useful. Although reported to have high performance in prediction tasks, NB classification is not the most optimal method for providing probability estimates of possible classes (Larsen, 2005) and is outperformed by other classifiers, *e.g.* logistic regression, when the size of training sample increases (Ng and Jordan, 2002). The exploratory analysis can as well be improved by considering more suitability variables, testing different spatial settings of the data model, and comparing *Phragmites* dynamics in different geographic regions. Finally, upon the availability of more historic data sets, running the CA model for more time steps allows better validation of its results as the behavior of the model might vary when more generations in the future are simulated.

ACKNOWLEDGMENTS

This work was partially conducted within the IBAM project (Integrated Bayesian risk analysis of ecosystem management in the Gulf of Finland), supported by the Baltic Organizations Network for Funding Science EEIG.

REFERENCES

Altartouri, A., and Jolma, A. (2012). Mining cellular automata rules: The use of a Naïve Bayes classifier to provide transition rules in Phragmites simulation. In: N.N. Pinto, J. Dourado & A. Natálio (eds), Proceedings of CAMUSS The International Symposium on Cellular Automata Modeling for Urban and Spatial Systems, pp. 79-90. Oporto, Portugal, November 8-10.

Bart, D., and J.M. Hartman (2003). The Role of Large Rhizome Dispersal and Low Salinity Windows in the Establishment of Common Reed, Phragmites australis in Salt Marshes: New Links to Human Activities. *Estuaries*, 26(2), 436-443.

Cheng, J., and R. Greiner (1999). Comparing Bayesian Network Classifiers. In: Proceedings of UAI '99 the Fifteenth Conference on Uncertainty in Artificial Intelligence, Stockholm, Morgan Kaufmann, San Francisco, CA, pp. 101-108.

Dormann, C.F., *et al.* (2007). Methods to account for spatial autocorrelation in the analysis of species distributional data: a review. *Ecography*, 30(5), 609-628.

Ekebom, J., P. Laihonen, and T. Suominen (2003). A GIS-based step-wise procedure for assessing physical exposure in fragmented archipelagos. *Estuarine, Coastal and Shelf Science*, 57, 887-898.

Elith, J., *et al.* (2006). Novel methods improve prediction of species' distributions from occurrence data. *Ecography*, 29, 129-151.

Fonstad, M.A. (2006). Cellular automata as analysis and synthesis engines at the geomorphology-ecology interface. *Geomorphology,* 77(3, 4), 217-234.

Hochachka, W.M., *et al.* (2007). Data-Mining Discovery of Pattern and Process in Ecological Systems. *Journal of Wildlife Management*, 71(7), 2427-2437.

Ikonen, I., and E. Hagelberg (eds.) (2007). Read Up on Reed! End report of the Reed Strategy -project (Interreg IIIA – programme). Southwest Finland Regional Environment Centre. pp.60.

King, R., W. Deluca, D. Whigham, and P. Marra (2007). Threshold Effects of Coastal Urbanization on *Phragmites australis* (Common Reed) Abundance and Foliar Nitrogen Chesapeake Bay. *Estuaries and Coasts*, 30(3), 469-481.

Larsen, K. (2005). Generalized Naive Bayes Classifiers. SIGKDD Explorations Newsletter, 7(1), 76-81.

Ménard A., and D.J. Marceau (2005). Exploration of spatial scale sensitivity in geographic cellular automata. *Environment and Planning B: Planning and Design*, 32(5), 693-714.

Minasny, B., A.B. McBratney (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. *Computers and Geosciences*, 32 (9), 1378-1388.

Mitchell, T.M. (2005). Generative and discriminative classifiers: Naive Bayes and logistic regression. In Machine Learning. http://www.cs.cmu.edu/~tom/mlbook/NBayesLogReg.pdf

Ng, A.Y., and M.I. Jordan (2001). On discriminative vs. generative classifiers: a comparison of logistic regression and naive Bayes. In: T.G. Dietterich, S. Becker, Z. Ghahramani (eds) NIPS. MIT Press, MA, pp 841-848.

Olden, J.D., J.J. Lawler, and N. LeRoy Poff (2008). Machine learning methods without tears: a primer for ecologists. *The Quarterly review of biology*, 83(2), 171-193.

Yang, Y., and G.I. Webb (2009). Discretization for naive-Bayes learning: managing discretization bias and variance. *Machine learning*, 74(1), 39-74.

Zimmermann, N.E., *et al.* (2010). New trends in species distribution modelling. *Ecography*, 33(6), 985-989.