**Not developed yet? Alternative ways to include locations without changes in land use
change models**

Tenley M. Conway* and Christopher Wellen.
Department of Geography, University of Toronto- Mississauga, Mississauga, Canada

*corresponding author: tenley.conway@utoronto.ca

Abstract

Biogeographers have developed a new generation of statistical models called presence-only models, which require no data concerning the absence of a species and do not assume that the absence of a species indicates habitat unsuitability. Both characteristics are especially useful when modeling a species which is actively spreading across a landscape. While urban expansion is sometimes equated to an invading species, the applicability of presence-only models have not yet been explored when modeling urban growth. This paper compares predictions of urban growth using a presence-only model (Ecological Niche Factor Analysis) and a more traditional presence-absence model (logistic regression). An additional model used pseudo-absence sites, from the presence-only model output, as input into the presence-absence model. The models were applied to New Jersey's Barnegat Bay watershed. Overall, the traditional presence-absence model performed the best, although the presence-only model was sufficiently similar to warrant further exploration of presence-only models when no reliable absence data (i.e. locations where no conversion occurred) exists. However due to data formatting requirements of the presence-only model, it is difficult to accommodate data pertaining to administrative boundaries, which are inherently Boolean. Finally, the output based on the pseudo-absence approach over-predicted urban conversion when compared to the other approaches.

Keywords: land use modeling; ecological niche factor analysis; logistic regression; urban development

1. Introduction

Most simulations of land use/ land cover (LULC) change rely on categorical data to calibrate and validate the model, with these data often having an unclear level of accuracy or error structure. Only recently has the issue of data misclassification, and procedures to use when incomplete error information is available, been explored within LULC models. For example, Pontius and Petrova (2010) developed a method to evaluate predicted results when the level of accuracy is unknown for the reference data. Another approach has been used in the habitat suitability modeling literature, where methods were developed to reduce the influence of data points associated with potentially high levels of uncertainty (Hirzel et al. 2002a). Traditional habitat modeling methods rely on species presence and absence data, but the absence data is often error prone as species absence is commonly assumed based on a lack of known presence (Chefaoui and Lobo 2008). The alternative methods instead focus only on locations where species have been observed (i.e. presence-only methods: Carpenter et al. 1993, Farber and Kadmon 2003, Guo et al. 2005, Hirzel et al. 2002b, Lobo et al. 2006).

The goal of habitat suitability modeling is parallel to many LULC models that focus on urban conversion: both seek to create suitability maps indicating likelihood of occurrence typically based on presence (location of species sighting or recent urban conversion) and absence (no species sighting or urban conversion) data. To date, the utility of the newer presence-only approaches in LULC modeling has not been examined.

This paper explores the value and applicability of alternative ways to specify sites with no known land conversion (i.e. absence data) in LULC modeling, by comparing a presence-only method (Ecological Niche Factor Analysis) to a standard presence-absence modeling method (logistic regression). Additionally, implications of using randomly chosen absence sites versus

those identified through a pseudo-absence selection method based on the output of the Ecological Niche Factor Analysis are examined. The study area is the urbanizing Barnegat Bay Watershed in coastal New Jersey (USA), with urban conversion as the focus of the model. The following sections describe presence-only models, including situations where they may be the best approach to reduce model error; detail the methods and results used in the case study; and provide a discussion of the broader implications of absence site inclusion methods.

2. Background

While habitat suitability models traditionally use binary presence-absence data (Periera and Itami 1991; Hirzel et al. 2002a; Giordano et al. 2010), it has long been recognized that the observed species absences may be misleading. Inaccuracies may be especially high for animals with large home ranges, or meaningless when modeling invading species (Hirzel et al. 2002a). Furthermore, traditional sources of data (i.e. atlases or museum collections) often include no attempt to determine where species are not located (Chefaoui and Lobo 2008).

Given these issues with accurately identifying species absence locations, several primary approaches have been used to generate appropriate absence data points (Chefaoui and Lobo 2008): (1) randomly choosing absence points from all areas where the species has not been found (Stockell and Peter 1999), (2) randomly selecting absence locations, but weighting them in light of verified absence locations (Zaniewski et al. 2002), (3) including absence points from a circular buffer around each presence point (Hirzel et al. 2002a), (4) including points with characteristics different from presence sites (Zaniewski et al. 2002), and (5) including points with characteristics similar to presence sites (Lütolf et al. 2006).

An alternative solution is an approach that does not require any absence data. To date, several of these methods have been proposed: Ecological Niche Factor Analysis (Hirzel et al. 2002a), Multi-Dimensional Niche Envelope (Farber and Kadmon 2003; Busby 1991; Lobo et al. 2006), DOMAIN (Carpenter et al. 1993), and Support Vector Machines (Guo et al. 2005). Additionally, some researchers have used presence-only methods to generate pseudo-absences that can then be included in standard presence-absence models (Wisz and Guisan 2009).

Previous studies have shown that presence-absence models usually produce better results when there is a high likelihood that absences are true absences (i.e. unoccupied sites with unsuitable conditions) and when there is a very high number of presence cells on the landscape—that is, all high and moderate suitability locations are occupied (Hirzel et al. 2002a). But biogeographers have found presence-only methods to be a better choice when no reliable absence data exists (Hirzel et al. 2002a), when the desire to avoid false-positives is high (Segurado and Araújo 2004), and when equilibrium assumptions are violated (Brotons et al. 2004). This last point refers to species that have recently invaded a new landscape and have not had sufficient time for all high suitability sites to be occupied. Thus, unoccupied sites do not necessarily mean unsuitable sites.

The meaning of absence (or no LULC change) sites in binary land use/ land cover models has received minimal attention. Most land cover models requiring absence data include all possible data points (Poelmans and Rompaey 2009; Hu and Lo 2007), use a random selection (Fragkias and Seto 2007; Mertens and Lambin 2000), or a spatially stratified random sampling approach (Batisani and Yarnal 2009), often with little discussion given to the potential for error associated with absence data (or for that matter, presence data).

The prevalence of logistic regression, and other binary presence-absence based methods, in land use change modeling indicates an acceptance of the assumptions around true-absences and equilibrium that these methods require, yet there is some ground to question them, particularly in urbanizing landscapes. In addition to the normal errors that classification of aerial and satellite imagery are prone to, an absence of urban change may not reflect unsuitability of that site. Similar to an invasive species, urban development may simply need more time for all suitable sites to become occupied. Disequilibrium may also occur when there has been a recent shift in the broad drivers of land use/land cover change (i.e. regional economics, cultural norms, etc.), increasing the demand for urban uses in a given region. Finally, there may be some situations where absence of known land conversion does not necessarily mean absence of actual land conversion. For example, presence-only methods may be a useful way to incorporate building permit data— which is often incomplete—, or model exurban landscapes with significant canopy cover— where the likelihood of not capturing isolated buildings from imagery is high— resulting in a large number of false positive absence points.

At the same time, the accuracy of absences, while a serious concern in habitat modeling, may be less of a concern when modeling LULC change, because in many cases absences likely have similar accuracy to the presence location when derived from the same input imagery and reference data. LULC modeling thus serves as an opportunity to test the accuracy of presence-only and presence-absence methods based on pseudo-absence data, as the results can be compared to results generated using relatively reliable absence data.

2.1 Ecological Niche Factor Analysis

One of the more common presence-only modeling techniques is the Ecological Niche Factor Analysis (ENFA), which combines concepts from ecological niche theory and factor analysis. Ecological niche theory defines a niche as a hyper-volume in the multidimensional space of ecological variables within which a species can maintain a viable population (Hutchinson 1957). Habitat suitability (or the likelihood of a viable population) decreases as one moves away from the optimum ecological conditions. The ENFA approach identifies the multidimensional space where a specific species can exist, creating habitat suitability output based on the similarity of each location to that set of optimum conditions.

More specifically, ENFA compares differences in the values of explanatory variables between presence locations and all data points for the study area. In this way, no assumptions are made about non-presence locations. In addition, ENFA differs from logistic regression in that it does not reject any input variable. In a step-wise selection processes, often used with logistic regression, causal variables can be removed from the model in favor of other, potentially spurious, correlations. Alternatively, ENFA weights variables, but all are retained.

Hirzel et al. (2002a) provide a detailed description of the calculations associated with ENFA. Through a factor analysis that includes all potential explanatory variables the first factor, called marginality, is generated by maximizing the distance between the centriods of the hyper-ellipsoids of the presence locations (species distribution) and all locations in the study site (the global distribution). The remaining factors, called specialization factors, maximize the variance of the global distribution and are perpendicular to the marginality factor and to each other. Thus the first factor describes where the presence sites are in relation to the entire study area (expressed in units of standard deviation), while the remaining factors indicate how broad a

range of values (for the given set of explanatory variables) is captured by the presence data (Hirzel et al. 2002a).

Once each factor is defined, habitat suitability maps can be created by selecting the number of factors to retain, usually using a crooked stick approach. A habitat suitability value for a given grid cell is then calculated separately for each factor by comparing the cell value to a histogram of the species presence cell values for that factor. The suitability is equal to the number of cells with values between the grid cell and the median value divided by the total cells with species presence. The suitabilities associated with each factor are summed using a weighted approach based on eigenvalues, with the marginality component weighted as one plus its eigenvalue. This usually places most of the weight on the marginality component (Hirzel et al. 2002a). The summed values are then rescaled, bounded between zero and one, with the combined factor output maps interpreted as estimates of habitat suitability.

Alternatively, ENFA-based suitability models have been combined with traditional presence-absence models. Locations with low values in the ENFA map are selected to represent pseudo-absence locations, based on the idea that these locations are less likely to be false absences (Wisz and Guisan 2009). These pseudo-absence data are then combined with presence data and incorporated into standard binary modeling approaches, like logistic regression. Studies that have used these combined models have mixed results. Some indicate that the approach does better than either ENFA or logistic regression-based models by themselves (Chefaoui and Lobo 2008; Engler et al. 2004; Jiménez-Valverde et al. 2008). Others based on simulated data indicate that simply selecting random non-presence locations gives better results (Wisz and Guisan 2009). This study is the first to use real data with mapped absences (i.e. no urban conversion) as a control when evaluating the use of pseudo-absences.

3. Study area and land use data

3.1 Barnegat Bay

The study area is the Barnegat Bay watershed, located on the outer coastal plain in southeastern New Jersey, USA (Figure 1). The watershed drains 1417 km² into the Barnegat Bay estuary. Oak-pine upland forests dominate the region, while freshwater wetlands are located throughout the study area and salt marshes are present along the edge of the estuary. The region has long been valued for its natural amenities, with tourism beginning as early as the 1700s on the barrier islands (Lloyd 1990). The number of summer tourists greatly increased with the completion of the railway lines from New York City and Philadelphia in the 1870s, and a second wave of tourism growth occurred in the 1940s and 1950s with the completion of two bridges from the mainland to the barrier islands and a highway (the Garden State Parkway) running along the New Jersey coast (Cunningham 1958).

Early development followed typical near-shore patterns, with most buildings located immediately landward of the shoreline (BBEP 2001). The year-round population was very low and inland urban development limited to a few small towns scattered throughout the region until in the middle of the 20th century (BBEP 2001). However, the rate of urban development exploded after World War II, associated with a population increase of more than 800% between 1950 and 2000 (US Census 1990; 2001). This period of growth has taken the form of large-scale suburban residential development located near the estuary and low density exurban development further in-land. Additionally, the barrier islands are now completely built out.

<Figure 1 about here>

The LULC datasets created by the NJDEP (2001) for 1986 and 1995 were used in the calibration process, while a 2002 NJDEP dataset was used to validate predictions (NJDEP 2007).

All three datasets were initially derived from digital orthophotographs using a head's up digitizing approach. The images were classified based on a modified Anderson et al. (1976) scheme. For the LULC model, the datasets were converted to raster images with a cell size of 60 m.

In 1986, the study area was 18% urban, increasing to 28% by 2002. However, the majority of land was still forest (42%) and wetlands (26%); 1% was used for agriculture and 3% was classified as barren land, a combination of transitional land, sand, and gravel mines (NJDEP 2001). Within the 2002 urban class, just over 10% was classified as rural residential, defined as residential property on parcels larger than 1 acre (0.4 ha), which are often surrounded by or adjacent to forest. As the minimum mapping unit for the dataset is 0.4 ha, approximately 1 acre, and many exurban building sites are surrounded by forest cover, it is likely that the total amount of rural residential land uses in the study area is underrepresented in the dataset. Thus, there is the potential for non-urban areas to contain more errors (false absences) than errors for classified urban locations (false presences).

In addition, the urban development rate increased in the area during the 1990s and 2000s, in part due to expansion pressure associated with the Atlantic City, Philadelphia, and New York City metropolitan areas, as well as people drawn to the area by the natural amenities of the region. Thus, it may be incorrect to assume that the study area is in a state of equilibrium during the time frame of the study. As a result, the use of a presence-only model may be most appropriate.

3.2 Explanatory Variables

To calibrate the models, we selected explanatory variables that represent accessibility to built and natural amenity features, neighborhood and site conditions, and spatial land policies (Table 1). These variables reflect factors that are commonly included in LULC models emphasizing urban conversion location (Berry et al. 1996; Verburg et al. 2004). The first four variables reflect measures of accessibility to built and natural amenities: proximity to protected open space (including federal, state, municipal and privately owned protected land); distance to urban land uses; driving distance to nearest highway exit and nearest barrier island bridge. This last variable represents a measure of driving distance to the beach from locations on the mainland.

Five aspects of site condition were examined. First, distance to the nearest road, a simple measure representing the ease of gaining access to the site during construction (Turner et al. 1996), was calculated. The second site variable was location inside or outside the floodplain, based on NOAA's 100 year flood line. While development is regulated adjacent to streams, wetlands, and other waterbodies, residential development is often still legally allowed within the broad floodplains in the watershed. Third, existing land cover is often an important locating factor of new urban development, so variables were specified to represent the existing land cover classes of forest, agriculture, wetlands, and barren land. Forth, a variable representing current access to sewer service was included. Fifth, the density of existing neighborhood urban uses and protected open space was included.

Finally, two variables representing spatial land use policies that constrain or enable development in certain areas were included in the analysis. The first reflects cells inside or outside the Pinelands Management Areas (Figure 1), which is regulated by the restrictive Pinelands Management Comprehensive Plan. The other policy variable included is a binary

representation of growth areas defined under the New Jersey Coastal Area Facility Review Act (CAFRA), which regulates almost all development activities in a zone that includes most land in the study area outside the Pinelands management areas.

Table 1 lists the dates of the available data representing the potential locating factors. In some cases, data corresponding to the dates of the land use datasets exist (i.e. distance to protected and open space), while only one point in time is available for other variables. Some variables do not change over the study period (i.e. Pinelands boundary), so having only a single dataset is not a problem. However, the variables related to features that have likely changed for which only one date in time is available may bias the model. For example, several locating variables are derived from a dataset representing road locations in 1995, the end date of the calibration time period. However, this bias should apply equally to the different modeling approaches, minimizing concern in this study.

<Table 1 about here>

4. Analytical methods

One suitability map was generated using ENFA methods, following Hirzel et al. (2002a), and six maps were made using logistic regression (Table 2). The model specifications were based on urban conversions from the calibration phase, 1986 to 1995, which were used to generate predicted suitability maps for the validation phase, 1995 to 2002. When possible, conditions reflecting the start of the time step (1986 or 1995) were used as the input variables. All analyses used a mask to identify areas that could undergo urban conversion: areas which were not already urban, within formally protected open spaces, and/or open water. All suitability maps were then evaluated using the area under the Relative Operating Characteristic curve (ROC-AUC), a

common statistic used to evaluate predicted suitability (Schneider and Pontius 2001), with differences between maps explored through visual comparison and correlation analysis of suitability histograms. The following sections discuss how the suitability maps were created using ENFA and logistic regression methods.

<Table 2 about here>

4.1 ENFA

Since ENFA requires all explanatory variables be normally distributed and continuous, the first step was converting Boolean data layers to distance layers, where the value of each cell was equal to the Euclidean distance to the nearest cell with a value of 1 (presence); and to density layers, where the value of each cell was equal to the proportion of '1' cells within a defined neighborhood. Both 1 km and 5 km square neighborhoods were considered. Variables were also transformed using square root or Box-Cox transformations, where appropriate. Finally, all variables were standardized.

Only one form of each variable was included into the analysis, if the variable was retained at all, with a trial and error approach used to determine the set that resulted in the highest marginality, or difference between conditions where urban change occurs and average conditions across the entire study area. The analysis was completed using Biomapper software (Hirzel et al. 2002b).

The factor analysis was then completed and the first six factors were retained, as indicated by the broken stick method. Suitability maps were then created for the calibration period using the medians algorithm described by Hirzel et al. (2002a), based on histograms of 250 classes and factor weights derived from the eigenvalues. Finally, the suitability map was rescaled with a linear transformation, so suitability could potentially range from 0 to 1.

To create the suitability map for the validation period, 1995 to 2002, the score matrix generated from ENFA during the calibration phase was applied to the 1995 data. Variables for which we had data for both the calibration and evaluation period were transformed in the validation period in exactly the same way as they were in the calibration period. No Box-Cox transformations were used with these variables, as they cannot be easily recreated with another dataset. The updated variables were also transformed with the means and standard deviations from the calibration period to assure that the same value of a given variable had exactly the same meaning between study periods. However, the histograms for the same variable from the two time periods tended to be quite similar, so the use of calibration means and standard deviations may not have been necessary. A suitability map was then created from the updated factors using the same methods as those employed during the calibration period. This process is based on the assumption that the variables that explained urban change in the 1986 to 1995 period exhibit exactly the same relationship with urban change during the 1995 to 2002 period, a common assumption is land use/land cover change models.

4.2 Logistic regression

Binary logistic regression quantifies the risk of a binary outcome (i.e. change to urban or no change to urban) using explanatory variables to differentiate cases where change did or did not occur. Six logistic regression-based models were created in this analysis, differing in the way absence locations and explanatory variables were selected (Table 2).

Binary logistic regression produces the most robust results when the number of values for the two categories is equal, regardless of the actual distribution (Hosmer and Lemeshow, 2000). Since there were many more absence (no change) sites than the 13,186 presence (urban change)

sites, a subset of absence locations was used. Three logistic regression models were created by randomly selecting absence cells, defined by the classified data. Specific locations were selected using the Sampling Design Tool from the U.S. National Oceanic and Atmospheric Administration's Biogeography Branch (Menza and Buja 2008). The Moran's I value of the layer of chosen absence locations is 0.0043, indicating nearly perfect random dispersion. The three other logistic regression models' absence samples were based on the least suitable cells for urban conversion identified in the ENFA-based suitability map.

Both absence samples were then applied to models based on three different sets of explanatory variables: (1) only variables retained in the final ENFA model, in their ENFA modified state (standardized, normally transformed, and distance or density specifications for Boolean variables); (2) only variables retained in the final ENFA model, in their original state; and (3) all the variables in their original form retained after a stepwise selection procedure. A value of 0.05 was used as a threshold for the Wald statistic for models created with variable set 3.

The six equations generated with these model specifications were used in Idrisi's Raster Calculator (Eastman 2006), updating input variables where appropriate, to create suitability maps for both the calibration and the validation periods.

5. Results

Table 3 provides the variables retained in the ENFA analysis, as well as their loadings on the factors that were retained when creating the suitability maps. Overall, the marginality, which reflects how different the locations of urban change are from the global average with respect to the explanatory variables, has a value of 0.966. Higher values of marginality indicate greater

difference between conditions of the average presence site and all locations (Hirzel et al. 2002a). More specifically, the mean value of the urban change locations is equal to 1.96 standard deviations away from the global mean. The specialization value (1.278) indicates that the presence variance is smaller than the global variance since it is above one. Thus, urban change occurs on a narrower range of conditions than found across available land in the study area. These results confirm the assumption that urban conversion locations are not random and can be identified by the locating factors.

<Table 3 about here>

The first factor of the ENFA model is highly negative correlated with distance to roads and sewers, and positively correlated with neighborhood urban density within a 5 km neighborhood. The results also indicate a tendency of urban conversions to take place closer to barrier island bridges but further from the Pinelands Management Areas than the study area average. To a lesser extent, urban conversions also occur in areas with a lower than average density of protected open space and higher than average density of CAFRA growth areas. Interestingly, both the distance to forest and the density of forest are negatively correlated with the marginality factor, indicating that urban conversions happen closer than average to forest, but in areas with lower forest density within a 5 km radius. This suggests that the fringe of the forest is more susceptible than elsewhere to urban conversion.

The six logistic regression-based model results show the greatest difference primarily between the two sampling scheme used to identify absences, with the models based on pseudo-absence sampling having much higher r^2 values than the random sample-based models (Table 4). This is not surprising, as the r^2 terms for all models are calculated only over the sampled data, not the entire study area, and the pseudo-absence sampling method only included no change sites

that were predicted to be most dissimilar from urban change locations. Using the original or transformed variables did not affect model fit significantly within the same absence sampling scheme.

<Table 4 about here>

In the three random sample models, some form of six explanatory variables appeared significant in all three models: distance to nearest highway exit, distance to local road, forest, agriculture, wetlands, and urban. When Boolean variables were allowed in the model, two to four additional variables with a binary distribution were included as significant variables in the models; when a step-wise selection approach was used two variables not included in the ENFA model were significant, with both representing binary factors (i.e. presence of barren land and floodplain).

The three pseudo-absence models had six significant variables in common: distance to highway exit, open space, distance to local road, wetlands, urban, and pinelands management area. Of these variables, the three that directly capture the urban fabric (distance to highway exit, distance to road, and urban) were also significant in the random sample-based models, highlighting the importance of these factors in the study area. Unlike the random sample models, there is not a disparity in significance for any explanatory variable when Boolean forms were allowed. As expected there were also no differences in the sign of coefficients for significant variables across the six models, although the relative size of coefficients did vary.

5.1 Comparison of urban change suitability surfaces – calibration period

Visual interpretation of the results yields some interesting findings (Figures 2 and 3). First, the surfaces generated using the pseudo-absence training data are quite saturated, especially outside

the areas where pseudo-absences were selected. Thus, pseudo-absence sampling led to over-prediction, as any raster cell not in the pseudo-absence sample was likely much more suitable for urban change, and so it was modeled as very likely to undergo change. Second, the three surfaces generated using the randomly selected training data are all quite different, though they all give a high importance to roads. The surface generated using the ENFA transformed data (dataset 1) in particular resembles the ENFA surface closely, suggesting that the differences in results between ENFA and logistic regression are partly due to the different ways data is typically formatted for these methods

<Figure 2 about here>

<Figure 3 about here>

Finally, the logistic regression surface generated using dataset 3 and the stepwise procedure yielded suitabilities lower than those of the ENFA generated surface, but did not disagree substantially regarding the areas of high suitability for urban change (i.e. areas along roads and near existing urban areas). On a regional scale though, these two surfaces do have different predictions regarding the suitability of inland areas, often near the Pinelands Management Area, for urban change. The ENFA surface predicts a greater likelihood of inland development than the logistic regression surface with dataset 3, possibly because the ENFA dataset does not include the Pinelands Management Area as a Boolean layer. Considering that the logistic regression surface calibrated with dataset 3 is the most likely urban change suitability surface to be produced through this method, the difference is noteworthy.

Similar to the suitability maps, histograms indicate a high degree of similarity with each other (Spearman rank correlation ≥ 0.74 for all pair-wise comparisons; Figure 4), but the distribution of values varied between models. An examination of the ENFA map and the logistic

regression with random sampling maps' histograms highlight the relatively high number of very low values and a general decrease in the number of cells moving from low to high values.

Alternatively, the logistic regression models that used pseudo-absences show a distinct bi-modal distribution, defined by very low and very high values, with more very high values predicted than very low ones. In fact, approximately half of the study area has a predicted suitability for urban change of 0.9 or higher, corresponding to roughly eight times the actual area of urban change between 1986 and 1995!

<Figure 4 about here>

The ROC-AUC statistics for the calibration period show that logistic regression maps based on the random sample have similarly high values, while the pseudo-absence maps have slightly lower values (Table 5). These results are similar to the relative performance other investigators found using maps generated from pseudo-absence sampling within logistic regression (Wisz and Guisan 2009). The ENFA map has the lowest ROC-AUC value, but all values are significantly different from the null model.

<Table 5 about here>

5.2 Comparison of urban change suitability surfaces – validation period

The surfaces generated using the models discussed above with the more recent data resemble the surfaces generated in the calibration phase (Figures 5 and 6). As in the calibration period, the saturation of the logistic regression maps based on pseudo-random sampling is apparent, while there is clear agreement between the ENFA map and the logistic regression maps generated with random sampling. As expected, the ROC-AUC statistics are somewhat lower than during the

calibration period (Table 5). Interestingly, there is less difference between the different sampling methods and variable forms in the evaluation period than in the calibration period.

<Figure 5 about here>

<Figure 6 about here>

6. Discussion

This study is the first to explore the presence-only models developed for habitat suitability modeling as a means of modeling urban conversions alone, or as part of a process to generate pseudo-absences in a presence-absence approach. Overall, the models generated with ENFA and logistic regression using random sampling of absences were relatively similar, though some differences were noted. The ENFA generated map had an ROC-AUC score consistently just below the maps of all logistic regression methods, indicating that when reliable absence data is lacking, ENFA could be used to model urban conversion. The ENFA methods open the possibility of manually identifying buildings/building sites on air photos or through field work, without needing to classify the entire study area. At the same time, the equilibrium assumptions of logistic regression did not appear to be violated to the point of rendering this method useless. More research is needed to determine how broadly this is the case, as very quickly growing areas may violate the equilibrium assumptions to a much greater extent than our study area.

However, differences between the presence-only and presence-absence methods influence the way the approaches are applied. Some of the variations in output are clearly a result of normalization and transformation of binary data, requirements of ENFA. For example, the logistic regression models using the modified variables assigned a lower importance to urban density than to forest or agriculture, while the logistic regression runs with the untransformed data gives a fairly high importance to urban density. Additionally, it was difficult to evaluate the

effect of the Pinelands Management Areas and the CAFRA growth areas in ENFA as both are inherently binary. Thus, a clear limitation of ENFA is its inability to directly incorporate Boolean variables.

Another challenge with ENFA implementation is the lack of an accepted process to identify the most parsimonious model. Unlike step-wise selection within logistic regression, a trial and error approach is usually used to select input variables into ENFA. While inclusion of variables that do not help differentiate presence sites should not negatively influence ENFA results, being able to identify the simplest set of variables—beyond a trial and error approach—with high predictive accuracy is valuable, particularly when considering data requirements for future predictions.

In line with biogeography applications comparing ENFA and logistic regression methods, we found using randomly selected absence locations a much better strategy than using pseudo-absences. It is worth noting that while the random sampling models tended to underestimate the risk of urban conversion relative to the pseudo-absence model, the normalizations and standardizations of variables tended to reduce this effect. There may, however, be some applications of suitability mapping, for example long-term land conservation planning, where the more saturated pseudo-absence predictions could be useful; that is, situations where the risk of under-predicting conversion is seen as worse than over-prediction.

Even with the bimodal predictions, the ROC-AUC values of the pseudo-absence models are nearly as high as the logistic regression models selected using random absences, and consistently higher than models generated with ENFA. These results are somewhat difficult to explain and likely reflect the weakness of the ROC-AUC statistic to detect this kind of problem. Indeed, Wisz and Guisan (2009) found that while the ROC-AUC statistic of suitability maps

generated using ENFA-selected pseudo-absences in their logistic regression model decreased if slightly more suitable locations were used as pseudo-absences, the ROC-AUC statistic never dropped below 0.9. Similarly, the high values of the Nagelkerke r^2 for the pseudo-absence models is misleading as this statistic only reflects the data used to calibrate the model. Given the two extremely different groups of cells, it is not surprising that the models would be very good at differentiating between them. It is also not surprising that these models would tend to estimate the risk of urban conversion of most cells as close to 1 or 0. Of course, these very high r^2 values also do not necessarily indicate that the models will be highly accurate in predicting different likelihoods of conversions outside the sample dataset.

7. Conclusion

This study was the first to compare presence-only methods with presence-absence methods in the context of land use\land cover change modeling. We found that the presence-only method ENFA modelled the risk of urban conversion nearly as well as the presence-absence methods when evaluated with the ROC-AUC. This indicates that the equilibrium assumptions of the presence-absence methods are not strongly violated in this study area, while the presence-only method may be a useful tool when absence data is incomplete. It is possible that in other study areas, particularly those characterized by fast growth or recent shifts in the drivers of growth, the equilibrium assumptions will be violated to the point that presence-only methods would prove more effective. Alternatively, the method of pseudo-absence selection employed in this study produced a nearly binary suitability map. While the ROC-AUC statistics for suitability maps generated with the pseudo-absence technique were reasonable, this approach clearly

overestimated the suitability for urban conversion, raising questions about the usefulness of such a technique.

Literature Cited

- Anderson, J. R., Hardy, E. E., Roach, J. T., and Witmer, R. E., 1976. *A land use and land cover classification system for use with remote sensor data*. Geological Survey professional paper 964. Washington, DC: US Government Printing Office.
- Batisani, N., and Yarnal, B., 2009. Urban expansion in Centre County, Pennsylvania: Spatial dynamics and landscape transformations. *Applied Geography*, 29 (2), 235-249.
- BBEP, 2001. *Barnegat Bay Estuary Program Comprehensive Conservation and Management Plan*. Toms River, NJ: Barnegat Bay Estuary Program.
- Berry, M.W., Hazen, B.C., MacIntyre, R.L., and Flamm, R.O., 1996. Lucas: a system for modeling land-use change. *IEEE Computational Science and Engineering*, 3(1), 24-35.
- Brotons, L., Thuiller, W., Araújo, M.B., and Hirzel, A.H, 2004. Presence-absence versus presence-only modeling methods for predicting bird habitat suitability. *Ecography*, 27, 437-448.
- Busby, J.R., 1991. BIOCLIM- a bioclimate analysis and prediction system. In: C.R. Margules and M.P. Austin, eds. *Nature Conservation: Cost Effective Biological Surveys and Data Analysis*. Melbourne, Australia: CIRSO, 64-68.
- Carpenter, G., Gillison, A.N., and Winter, J., 1993. DOMAIN: a flexible modeling procedure for mapping potential distributions of plants and animals. *Biodiversity and Conservation*, 2, 667-680.
- Chefaoui, R.M. and Lobo, J.M., 2008. Assessing the effect of pseudo-absences on predictive distribution model performance. *Ecological Modelling*, 210, 478-486.

Cunningham, J.T., 1958. *The New Jersey Shore*. New Brunswick, NJ: Rutgers University Press.

Eastman, R., 2006. IDRISI Rel. 15. Worcester: Clark Labs, Inc.

Engler, R., Guisan, A., and Rechsteiner, L., 2004. An improved approach for predicting the distribution of rare and endangered species from occurrence and pseudo-absence data. *Journal of Applied Ecology*, 41, 263-274.

Farber, O. and Kadmon, R., 2003. Assessment of alternative approaches for bioclimatic modeling with special emphasis on the M-distance. *Ecological Modelling*, 160, 115–130.

Fragkias, M. and Seto, K.C., 2007. Modeling urban growth in data-sparse environments: a new approach. *Environment and Planning B: Planning and Design*, 34(5), 858-883.

Giordano, P. F., Navarro, J. L., and Martella, M.B., 2010. Building large-scale spatially explicit models to predict the distribution of suitable habitat patches for the Greater rhea (*Rhea americana*), a near-threatened species. *Biological Conservation*, 143(2), 357-365.

Guo, Q.H., Kelly, M., and Graham, C.H., 2005. Support vector machines for predicting distribution of sudden oak death in California. *Ecological Modelling*, 182(1), 75-90.

Hirzel, A.H., Hausser, J., Chessel D., and Perrin, N., 2002a. Ecological-niche factor analysis: How to compute habitat-suitability maps without absence data? *Ecology*, 83(7), 2027-2036.

Hirzel, A.H., Hausser, J., and Perrin, N., 2002b. Biomapper 4.0 Lusanne, Lab for Conservation Biology. URL: <http://www.unil.ch/biomapper>.

Hosmer, D.W. and Lemeshow, S., 2000. *Applied Logistic Regression*. 2nd Edition. New York: John Wiley and Sons, Inc.

Hu, Z. and Lo, C.P., 2007. Modeling urban growth in Atlanta using logistic regression. *Computers, Environment and Urban Systems*, 31(6), 667-688.

- Hutchinson, G.E., 1957. Concluding remarks. *Cold Spring Harbour Symposium on Quantitative Biology*, 22, 415-427.
- Jiménez-Valverde, A., Gómez, J.F., Lobo, J.M., Baselga, A., Hortal, J., 2008. Challenging species distribution models: The case of *Maculinea nausithous* in the Iberian Peninsula. *Annales Zoologici Fennici*, 45(3), 200-210.
- Lloyd, J.B., 1990. *Six miles at sea: A pictorial history of Long Beach Island*, Harvey Cedars, NJ: Down the Shore Publishing and Sand Paper, Inc.
- Lobo, J.M., Verdu, J.R. and Numa, C., 2006. Environmental and geographical factors affecting the Iberian distribution of flightless *Jekelius* species (Coleoptera: Geotrupidae). *Diversity and Distribution*, 12, 179-188.
- Lütolf, M., Kienast, F., and Guisan, A., 2006. The ghost of past species occurrence: improving species distribution models for presence-only data. *Journal of Applied Ecology*, 43, 802-815.
- Menza, C. and Buja, K., 2008. *Sampling Design Tool for ArcGIS*, National Atmospheric and Oceanic Administration Center for Coastal Monitoring and Assessment [online]. Available from: <http://ccma.nos.noaa.gov/products/biogeography/sampling/welcome.html> [Accessed 10 February 2009].
- Mertens, B. and Lambin, E.F., 2000. Land-cover-change trajectories in Southern Cameroon *Annals of the Association of American Geographers*, 90 (3), 467-494.
- NJDEP, 2001. *NJDEP 1995/97 land use/land cover update* [online]. Available from: <http://www.state.nj.us/dep/gis/> [Accessed 10 April 2008].

NJDEP, 2007. *NJDEP 2002 land use/land cover* [online]. Available from:

<http://www.state.nj.us/dep/gis/> [Accessed 10 April 2008].

Periera, J.M.C. and Itami, R.M., 1991. GIS-based habitat modeling using logistic multiple regression: a study of the Mt. Graham Red Squirrel. *Photogrammetric Engineering and Remote Sensing*, 57, 1475-1486.

Poelmans, L. and Van Rompaey, A., 2009. Detecting and modelling spatial patterns of urban sprawl in highly fragmented areas: A case study in the Flanders-Brussels region. *Landscape and Urban Planning*, 93(1), 10-19.

Pontius Jr, R.G. and Petrova, S., 2010. Assessing a predictive model of land change using uncertain data. *Environmental Modeling and Software*, 25(3), 299-309.

Segurado, P. and Araújo, M.B., 2004. An evaluation of methods for modeling species distributions. *Journal of Biogeography*, 31, 1555-1568.

Schneider, L.C. and Pontius, R.G. Jr, 2001. Modeling land-use change in the Ipswich watershed, Massachusetts, USA. *Agriculture, Ecosystems and Environment*, 85, 83-94.

Stockwell, D.R.B. and Peters, D.P., 1999. The GARP modeling system: problems and solutions to automated spatial prediction. *International Journal of Geographic Information Systems*, 13, 143-158.

Turner, M.G., Wear, D.N., and Flamm, R.O., 1996. Land ownership and land-cover change in the southern Appalachian Highlands and the Olympic Peninsula. *Ecological Applications*, 6(4), 1150-1172.

US Census, 1990. *Census of Population and Housing, Historical Population Counts and STF-1A (Profile 1)*, Washington, DC: US Census Bureau.

US Census, 2001. *US Census 2000, General Demographic Profiles* Washington, DC: US Census Bureau.

Verburg, P.H., Ritsema van Eck, J.R., de Nijs, T.C.M., Dijst, M.J., and Schot, P., 2004.

Determinants of land-use change patterns in the Netherlands. *Environment and Planning B: Planning and Design*, 31, 125-150.

Wisz, M. and Guisan, A., 2009. Do pseudo-absence selection strategies influence species distribution models and their predictions? An information-theoretic approach based on simulated data. *BMC Ecology*, 9(1), 8.

Zaniewski, A.E., Lehman, A., and Overton, J., 2002. Predicting species spatial distributions using presence-only data: a case study of native New Zealand ferns. *Ecological Modelling*, 157, 261-280.

Table 1: Potential locating factors.

Category	Locating Factor	Dates Available	Units
Built and Natural Amenity Access	Distance to Protected Open Space	1986 and 1995	km
	Distance to Urban Land	1986 and 1995	km
	Distance to Highway Exits	1995	km
	Driving distance to Barrier Island Bridge	1995	km
	Distance to Road	1995	km
Neighborhood and Site Conditions	Floodplain (100 Year)	1995	binary
	Forest	1986 and 1995	binary
	Agriculture	1986 and 1995	binary
	Wetlands	1986 and 1995	binary
	Barren Land	1986 and 1995	binary
	Sewer Service	1995	binary
	Density of Protected Open Space	1986 and 1995	percent
	Density of Urban Land	1986 and 1995	percent
	CAFRA	1995	binary
Land Policy	Pinelands Management Areas	1995	binary

Table 2: The six logistic regression-based model specifications examined.

Variable Set	Variable form	Absence Sampling	Model Algorithm
Set 1: Limited to ENFA vars.	Normalized; Binary vars. transformed	Random	Full model
Set 2: Limited to ENFA vars.	Original format	Random	Full model
Set 3: All vars.	Original format	Random	Step-wise selection
Set 1: Limited to ENFA vars.	Normalized; Binary vars. transformed	Pseudo-Random	Full model
Set 2: Limited to ENFA vars.	Original format	Pseudo-Random	Full model
Set 3: All vars.	Original format	Pseudo-Random	Step-wise selection

Table 3: ENFA factor results.

Locating Factor	Variable		Correlation with factor					
	Transformation	Normalization	1	2	3	4	5	6
Roads	Distance	Box-Cox	-0.61	0.09	-0.25	0.12	-0.08	-0.07
Urban	Density (5 km)	Square Root	0.40	-0.22	0.05	-0.34	-0.14	0.67
Sewers	Distance	Square Root	-0.30	-0.73	0.56	-0.47	-0.13	0.45
Pinelands Management Areas	Distance	Square Root	0.27	-0.09	0.17	-0.20	0.66	-0.15
Driving distance to Barrier Island Bridge	Driving Distance	Square Root	-0.26	0.12	-0.06	-0.08	-0.02	0.01
Highway Exit – Driving Distance	Distance	Box-Cox	-0.25	-0.10	-0.31	0.02	0.59	0.30
Open Land	Density (5 km)	Square Root	-0.19	0.28	0.67	0.23	0.20	0.13
Agriculture	Distance	Square Root	-0.18	-0.06	0.15	-0.19	0.08	-0.31
Wetlands	Distance	None	0.18	0.04	0.15	-0.07	0.16	-0.11
CAFRA	Density (1 km)	Square Root	0.17	-0.05	-0.08	-0.05	-0.01	0.00
Forest	Density (5 km)	None	-0.15	-0.34	-0.05	-0.30	0.32	0.29
Forest	Distance	Square Root	-0.14	0.42	0.06	-0.64	-0.05	0.19

Table 4: Logistic regression beta coefficients.

Variable	Random Absence Sample			Pseudo-Absence Sample		
	Set 1	Set 2	Set 3	Set 1	Set 2	Set 3
Open land- binary	x	x	x	x	x	x
Open Land-denisty(5 km)	-0.28	-1.69	x	-1.17	-8.38	x
Open Land-distance	x	x		x	x	3.34
Urban-distance	x	x	-0.84	x	x	-3.49
Urban -denisty(1 km)	x	x	x	x	x	x
Urban -denisty(5 km)	0.22	0.86	x	1.69	13.92	x
Highway Exit - Distance	-0.21	-0.04	-0.05	-0.30	-0.17	-0.21
Driving distance to Barrier Island Bridge	-0.18	0.00	nr	-0.43	-0.02	nr
Roads-distance	-1.42	-9.50	-9.48	-2.95	-9.31	-11.34
Roads-denisty(1 km)	x	x	x	x	x	x
Roads-denisty(5 km)	x	x	x	x	x	x
Floodplain- binary	x	x	-0.65	x	x	nr
Floodplain (100 Year)-denisty(5 km)	x	x	x	x	x	x
Floodplain (100 Year)-distance	x	x	x	x	x	x
Forest- binary	x	0.33	0.86	x	1.42	nr
Forest-denisty(5 km)	-0.29	x	x	0.05	x	x
Forest-distance	-0.58	x	x	-0.98	x	x
Agriculture-binary	x	0.70	1.16	x	2.68	nr
Agriculture-distance	-0.32	x	x	-1.27	x	x
Agriculture-denisty(1 km)	x	x	x	x	x	x
Agriculture-denisty(5 km)	x	x	x	x	x	x
Wetlands-binary	x	-1.04	-0.34	x	-1.66	-2.66
Wetlands -denisty(5 km)	x	x	x	x	x	x
Wetlands-distance	0.20	x	x	0.46	x	x
Barren-binary	x	x	1.42	x	x	nr
Sewers- binary	x	0.41	0.61	x	-0.05	1.64
Sewers-distance	-0.03	x	x	-0.16	x	x
CAFRA-binary	x	0.04	nr	x	0.11	nr
CAFRA-denisty(1 km)	0.02	x	x	-0.39	x	x
Pinelands-binary	x	-0.49	-0.57	x	-1.89	-1.95
Pinelands Management Areas-distance	0.03	x	x	2.03	x	x
Nagelkerke R2	0.53	0.54	0.54	0.97	0.97	0.96

notes: x = not included in analysis; nr = not retained in stepwise analysis; bold = significant at $p < 0.05$

Table 5: ROC-AUC statistics for the various models.

Model	ROC-AUC Value	
	1986 - 1995	1995 - 2002
ENFA	0.767	0.727
Random Sample - Transformed Data	0.871	0.765
Random Sample - Original Data	0.873	0.767
Random Sample - Stepwise	0.835	0.760
Pseudo-absence Sample - Transformed Data	0.841	0.765
Pseudo-absence Sample - Original Data	0.831	0.768
Pseudo-absence Sample - Stepwise	0.795	0.756

Figures

Figure 1: The study area.

Figure 2: Predicted ENFA-based suitability for 1986-1995.

Figure 3: Predicted logistic regression-based suitability for 1986-1995 for (a) variable set 1 with random sampling, (b) variable set 2 with random sampling, (c) variable set 3 with random sampling, (d) variable set 1 with pseudo-absences, (e) variable set 2 with pseudo-absences, and (f) variable set 3 with pseudo-absences.

Figure 4: Histograms showing suitability of change 1986-1995 for logistic-regression-based suitability using (a) variable set 1 with random sampling, (b) variable set 2 with random sampling, (c) variable set 3 with random sampling, (d) variable set 1 with pseudo-absences, (e) variable set 2 with pseudo-absences, and (f) variable set 3 with pseudo-absences, and (g) ENFA-based surface.

Figure 5: Predicted ENFA-based suitability for 1995-2002.

Figure 6: Predicted logistic regression-based suitability for 1995-2002 for (a) variable set 1 with random sampling, (b) variable set 2 with random sampling, (c) variable set 3 with random sampling, (d) variable set 1 with pseudo-absences, (e) variable set 2 with pseudo-absences, and (f) variable set 3 with pseudo-absences.

Value

0 10

Kilometers

a.

b.

c.

d.

e.

f.

0 30
Kilometers

Value

High : 1

Low : 0

Value

0 10

Kilometers

a.

b.

c.

d.

e.

f.

0 30
Kilometers

Value

High : 1

Low : 0

N