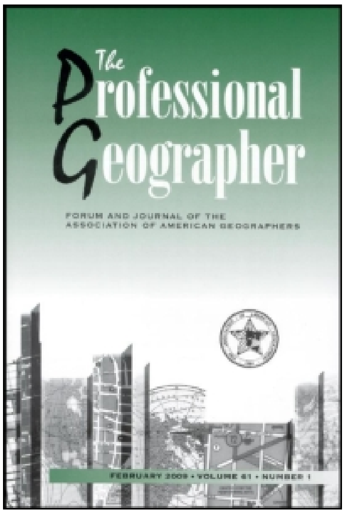

The Professional Geographer
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t788352615

The Influence of the “Mixed Pixel” Problem on the Detection of Analogous Forest Communities Between Presettlement and Present in Western New York

Barry J. Kronenfeld [a]; Yi-Chen Wang [b];Chris P. S. Larsen [c]
[a] George Mason University, [b] National University of Singapore, [c] University at Buffalo-State University of New York,

First published on: 12 April 2010

To cite this Article Kronenfeld, Barry J. , Wang, Yi-Chen andLarsen, Chris P. S.(2010) 'The Influence of the “Mixed Pixel” Problem on the Detection of Analogous Forest Communities Between Presettlement and Present in Western New York', The Professional Geographer, 62: 2, 182 — 196, First published on: 12 April 2010 (iFirst)

To link to this Article: DOI: 10.1080/00330120903546759
URL: http://dx.doi.org/10.1080/00330120903546759

PLEASE SCROLL DOWN FOR ARTICLE

The Influence of the "Mixed Pixel" Problem on the Detection of Analogous Forest Communities Between Presettlement and Present in Western New York*

Barry J. Kronenfeld
George Mason University

Yi-Chen Wang
National University of Singapore

Chris P. S. Larsen
University at Buffalo–State University of New York

We conducted a land change analysis to determine if the forest communities of presettlement and present contain areas that are analogous in composition, using surveys of the Holland Land Company (1797–1799) and U.S. Forest Service (1991–1993) from western New York. Gridded forest-type maps are produced from each survey using two models: a uniform model that assumes each data grid cell is occupied by a single forest type, and a mixture model in which grid cells are assumed to be occupied by multiple forest types in different proportions. The mixture model consistently detects a larger area of analogous communities in the two time periods at both global and local scales. **Key Words: analog communities, forest community change analysis, linear unmixing, presettlement land survey record, western New York.**

我们利用荷兰土地公司（1797 年至 1799 年）和美国林务局（1991 年至 1993 年）的数据，对纽约西部地区进行了土地用途变更的分析，以确定早期移民定居时期和现在的森林群落在组成上是否类似。对每组调查数据，我们使用了两种模型来生成森林类型的栅格图像：一种是统一模型，假设每个数据网格单元是单一的一种森林类型；另一种是混合模型，网格单元被假定为多个不同的森林类型按比例混合。在全局和局部尺度上，该混合模型可以稳定地检测到两个时期森林群落的相似性。**关键词：模拟社区，森林群落的变化分析，线性元分解，早期移民定居的土地调查记录，纽约州西部。**

Llevamos a cabo un análisis de la transformación de la tierra para determinar si las comunidades forestales de antes del asentamiento humano y en el presente contienen áreas que sean análogas en composición, utilizando estudios hechos en el occidente de Nueva York por la Compañía Holland Land (1797–1799) y los del Servicio de Bosques de los EE.UU (1991–1993). A partir de cada uno de esos estudios, se generaron mapas reticulados tipo forestal mediante la aplicación de dos modelos: un modelo uniforme en el que se asume que cada celda de la rejilla de datos es ocupada por un solo tipo de bosque, y un modelo mixto en el cual las celdas de la rejilla se asumen ocupadas por múltiples tipos de bosque, en diferentes proporciones. El modelo mixto consistentemente detecta un área más grande de comunidades análogas en dos períodos de tiempo a escalas tanto globales como locales. **Palabras clave: comunidades análogas, análisis de cambio en comunidades forestales, desmezclado linear, estudio de tierras prístinas, occidente de Nueva York.**

In analysis of remote sensing imagery, it is widely recognized that a single pixel may contain multiple land cover categories. Failure to account for these so-called mixed pixels

*The authors would like to acknowledge support from NUS funding # R-109-000-060-012/133 and thank the three anonymous reviewers for their insightful comments on the article.

The Professional Geographer, 62(2) 2010, pages 182–196 © Copyright 2010 by Association of American Geographers.
Initial submission, June 2008; revised submissions, March and June 2009; final acceptance, July 2009.
Published by Taylor & Francis Group, LLC.

increases classification uncertainty and results in bias against small land cover categories (Foody 1996). In change detection analyses, ignoring mixed pixels tends to result in erroneous estimation of the amount of disagreement between two images (Power, Sims, and White 2001; van Oort 2005). To avoid these biases, techniques have been developed to "unmix" each pixel into its source components using linear, stochastic, optical, geometric, or fuzzy mixture models (Ichoku and Karnieli 1996).

The problem of mixed pixels is ultimately caused by coarse data resolution and thus applies to other data contexts besides remote sensing. One such context arises when data collected at point locations are aggregated or interpolated either to increase the sample size for statistical estimation or to create a spatial grid for mapping purposes. The resulting mixture of data within each grid cell is analogous to the mixture of reflectances from multiple objects present in a single remote sensing pixel.

This study examines the effect of the mixed pixel problem on interpretation of historical changes that have occurred in the forests of the eastern United States following European settlement. At this spatiotemporal scale, our understanding has been greatly facilitated by the existence of presettlement land survey records (PLSRs), which contain records of the individual trees marked by the first European surveyors of frontier land throughout most of the United States. PLSRs have been used since the 1920s to reconstruct forest composition prior to major European settlement and have been compared with modern forest inventories to analyze changes that have occurred in the intervening centuries (Wang 2005). However, researchers are constrained by the sampling intensity of the original surveys. Although there exists substantial variation, a typical survey consisted of two to four trees marked at each survey corner, with corners spaced at half-mile intervals along survey lines running along the boundaries of 1 × 1 square mile township sections. Modern forest inventories are also limited in their spatial resolution, and although more intensive sampling of the present-day forest is possible, costs are generally prohibitive. The wide areas covered by these data sources, their limited sampling intensity, and lack of spatial correspondence have led several authors to aggregate or interpolate data to form gridded data representations with cell sizes ranging from 1 × 1 mile (He et al. 2000) to 6 × 6 miles (Friedman and Reich 2005).

One problem encountered in comparative PLSR studies is that few analogs to presettlement forest community types have been found in modern forests (Foster, Motzkin, and Slater 1998; Whitney and DeCant 2001; Friedman and Reich 2005). This lack of analog communities has caused several authors to limit the scope of analysis to certain forest types only (Batek et al. 1999; Radeloff et al. 1999). Most studies do not perform community change analysis at all; instead, research is limited either to the distribution of historical forest types (e.g., Leitner et al. 1991; Abrams and McCay 1996; Brown 1998; Black, Foster, and Abrams 2002; Cogbill, Burk, and Motzkin 2002; Bolliger et al. 2004; Bolliger and Mladenoff 2005; Wang 2007) or to changes in the distributions of individual taxa rather than communities (e.g., Cowell 1998; Radeloff et al. 1999; Dyer 2001). A few studies that primarily focus on changes in taxonomic composition also discuss (Whitney and DeCant 2003) or map (Foster, Motzkin, and Slater 1998; Friedman and Reich 2005) forest communities for both time periods but do not analyze changes in the location or spatial pattern of individual community types. By and large, presettlement research has either been confined to one time period or else focused on taxa rather than communities.

The perceived lack of modern analogs to presettlement forest communities has been attributed to the large differences in taxonomic composition observed between presettlement and present (Cowell 1998). This explanation cannot be complete, however, because the net changes observed in forest community types have been larger than those observed at the level of individual taxa (e.g., Foster, Motzkin, and Slater 1998; Friedman and Reich 2005). An ecological explanation for the greater change in community types is that there have been changes in stand-scale associations between taxa. An alternative explanation might be that data aggregation and coarse resolution analysis have mixed together existing analog communities. To distinguish between these two possibilities, it is necessary to consider the mixed pixel problem and apply appropriate unmixing techniques.

In this study, the distribution of forest community types in western New York in 1797–

1799, when the area was surveyed by the Holland Land Company (HLC), is compared to that in 1991–1993 when the area was surveyed by the U.S. Forest Service as part of the Forest Inventory and Analysis (FIA) program. To determine the effect of the mixed pixel problem on the discovery of analog communities, analysis is conducted via two models: a *uniform model*, in which each grid cell is assumed to be occupied uniformly by a single forest type, and a *mixture model*, in which each grid cell is posited to contain a mixture of different forest types. A linear unmixing technique is used to estimate forest type proportions within each grid cell under the mixture model.

Although the problem of coarse data resolution has been noted in other presettlement studies (e.g., Dyer 2001), no previous study has applied a mixture model to PLSRs. Several authors have used fuzzy classification, which is similar to linear unmixing, to represent boundary uncertainty and continuous gradients between presettlement communities (Brown 1998; Bolliger and Mladenoff 2005). These studies, however, did not interpret fuzzy membership values as proportions, nor did they address the implications of the mixed pixel problem on community change analysis. By applying an explicit mixture model to PLSRs, we aim to determine to what degree perceived changes in forest type distributions might be the result of coarse resolution analysis. In addition, this article illustrates a common scenario in which mixture problems occur and demonstrates how unmixing techniques can be applied in this scenario.

Linear Unmixing

The problem of estimating proportions of constituent components from mixed data is common to a variety of disciplines, ranging from chemistry to geology and air pollution studies (Akerjord and Christophersen 1996). In remote sensing, linear unmixing has often been applied to the mixed pixel problem, which arises when the spectral attributes of a single pixel result from the combined reflectance of multiple surface types. Linear unmixing can be considered a type of fuzzy classification; Nascimento, Mirkin, and Moura-Pires (2003) define a "proportional membership model" of fuzzy classification, and researchers in climate (McBratney and Moore 1985), soil science (Zhu 1997), vegetation mapping (Kronenfeld 2005), and geographic information systems (Burrough and McDonnell 1998) have all suggested a linear mixture be used to reconstruct underlying attributes from fuzzy membership values. However, Ichoku and Karnieli (1996) describe the "fuzzy" model as a distinct alternative to linear unmixing in the estimation of subpixel land cover in remote sensing applications. To avoid confusion, the term *linear unmixing* is used throughout this article.

The conceptual basis behind linear unmixing in this study is illustrated in Figure 1. We assume that homogenous forest communities of various types are distributed across the landscape in discrete patches (Figure 1A). Recorded data are sparse, and so data on taxonomic composition are interpolated to a regular grid (Figure 1B). Estimated taxonomic composition in

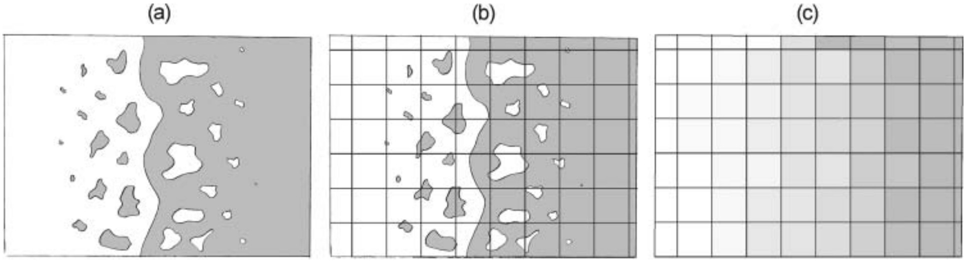

Figure 1 *Schematic diagram showing coarse resolution data derived from a patchwork of two forest communities along an environmental gradient (A). These patches occur at or finer than the resolution of the analysis grid (B). Although subpixel spatial distribution cannot be discerned, the proportion of each grid cell occupied by each forest type can be estimated, as indicated by shading in (C).*

each cell is a function of the proportion of the cell occupied by each forest type, plus an error component introduced by the interpolation process. Although it is impossible to reconstruct the underlying spatial pattern, forest type proportions within each grid cell (indicated by the shades of gray in Figure 1C) can be predicted from this taxonomic composition.

Numerically, linear unmixing seeks to express the observed taxonomic composition of each grid cell as the weighted average of a set of component forest types:

$$x_{ij} \approx \hat{x}_{ij} \equiv \sum_{q=1}^{k} p_{iq} c_{qj} i = 1 \ldots j = 1 \ldots m \tag{1}$$

where m = number of taxa; n = number of grid cells; k = number of forest types; x_{ij}= actual relative abundance of the jth taxon in the ith grid cell; $\hat{x}_{ij}$= modeled relative abundance of the jth taxon in the ith grid cell; p_{iq}= proportion of forest community type q in grid cell i; and c_{qj} = typical relative abundance of taxon j in forest community type q.

This formula can be represented succinctly in matrix form as:

$$\boldsymbol{X} \approx \hat{\boldsymbol{X}} \equiv \boldsymbol{PC}' \tag{2}$$

where $\boldsymbol{X}$ and $\hat{\boldsymbol{X}}$ are $n \times m$ matrices of actual and modeled species-relative abundances in each grid cell, $\boldsymbol{P}$ is an $n \times k$ matrix of forest type proportions in each grid cell, $\boldsymbol{C}$ is an $m \times k$ matrix of prototypical taxon abundances in each forest type, and C' is the transpose of C. The proportion matrix $\boldsymbol{P}$ is constrained by two conditions

$$0 \leq p_{iq} \leq 1, \; i = 1 \ldots n, \; q = 1 \ldots k \tag{3}$$

$$\sum_{q=1}^{k} p_{iq} = 1, \; i = 1 \ldots n \tag{4}$$

that denote that proportions must be nonnegative and sum to one.

Because the matrix product $\boldsymbol{PC}'$ models each element in $\boldsymbol{X}$ as a linear combination of constituent types, it is referred to as a *linear mixing model* (Akerjord and Christophersen 1996). A good model minimizes the difference between $\boldsymbol{X}$ and $\boldsymbol{PC'}$, usually in a least-squares sense. Commonly, the variance-weighted average squared difference of matrix elements is minimized, resulting in component types being located in principal component space (Akerjord and Christophersen 1996). Subtracting this difference metric from unity yields the percentage of the total variance of $\boldsymbol{X}$ explained by the mixture model; that is, the percentage variance explained (*PVE*) of principal components analysis (Kronenfeld 2005).

No unique solution to linear unmixing maximizes *PVE*, because for any model $\{\boldsymbol{P}, \boldsymbol{C}\}$, an equivalent model $\{\boldsymbol{P}_2, \boldsymbol{C}_2\}$ can be derived through simple algebraic transformation (Wolbers and Stahel 2005). To overcome this problem, $\boldsymbol{C}$ is often determined either from prior knowledge or by indirect analysis in which optimal constituent types are sought that conform to assumptions regarding the source domain. The most common assumption is that mixing is limited at least in some observations, so that constituent types can be found among the data points. Building on this assumption, several metrics have been developed to indicate the compactness and separation of data around component types (e.g., Fukuyama and Sugeno 1989; Gath and Geva 1989; Xie and Beni 1991). However, the coarse resolution of PLSRs makes it difficult to accept that the mixture of forest types within a grid cell would be limited. Therefore, we do not use these metrics in this article, nor do we presume to be able to identify the forest types that existed in either time period. Instead, we apply strictly the same classification method to both presettlement and modern forest inventory data at a coarse resolution and then compare the results of the uniform and mixture models. In this way, we seek to establish the degree to which discovery of analog communities is affected by the mixed pixel problem per se.

Once the forest type matrix $\boldsymbol{C}$ has been determined, each data point can be "unmixed" according to one of several mixture models (Ichoku and Karnieli 1996). A linear model is appropriate when observations are simple compositional mixtures of constituent components, as is the case with areal aggregations of internally homogeneous forest communities. Linear unmixing is achieved by projecting the vector representing the taxonomic composition of each grid cell onto the $(k - 1)$ dimensional

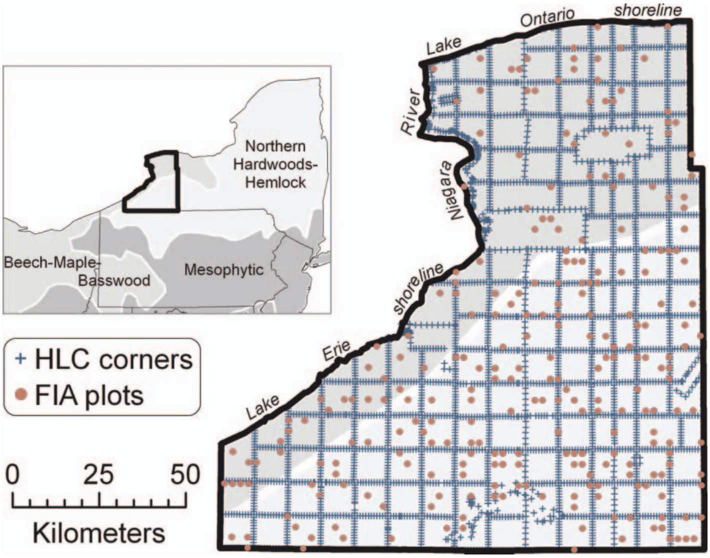

Figure 2 *Area of the Holland Land Company (HLC) survey in western New York, with sampling locations. Dyer's (2006) forest regions are shown in shades of gray and labeled in the inset. FIA = Forest Inventory and Analysis.*

hyperplane defined by the k component forest types. Given $\boldsymbol{C}$, linear unmixing produces a unique proportion matrix $\boldsymbol{P}$. Linear mixing is implemented in most remote sensing software, including ENVI (ITT Visual Information Solutions).

Study Area and Data

The study area is comprised of 162 townships surveyed by the HLC between 1797 and 1799 in western New York (Figure 2). These townships lie between the Pennsylvania state line to the south and Lake Ontario to the north and are bounded on the west by Lake Erie and the Niagara River. The area encompasses all or part of eight counties and covers approximately 14,400 km^2, extending across two commonly recognized physiographic sections. The northern part of the study area is located in the Erie Ontario Lowland, a section with relatively low, flat topography; the southern part belongs to the Appalachian Upland and has topography of dissected uplands, formerly glaciated in most areas (Fenneman 1938). These physiographic sections correspond closely to broad ecological (Bailey 1995) and vegetation (Dyer 2006) regions. The presettlement vegetation was dominated by beech (*Fagus grandifolia*) and sugar maple (*Acer saccharum*; Wang 2007).

The HLC township perimeter survey records were used to create maps of forest communities as they existed prior to major European settlement. Influenced by the Land Ordinance of 1785, the legal origin of the rectangular system of public land surveys (Wyckoff 1988), the private HLC developer employed a regular township survey system in western New York. Land was divided mostly into 6 × 6 mile townships (1 mile ≈ 1.609 km), but sizes of 4 × 6 and 7 × 6 miles were also used (Figure 2). Unlike the public land surveys for which the finer section-level data at 1 × 1 mile are available, the finer level data are not available for all the HLC townships and hence only data from the township perimeter surveys were used in this study. Posts were erected at half-mile intervals along the township perimeter survey lines, and neighboring trees, known as *bearing trees*, were blazed and inscribed to mark the location of each post. Species name, distance, and direction from each bearing tree to the post were recorded by the surveyors. These data were transcribed from microfilms of surveyors' manuscripts obtained from the HLC Archives at SUNY Fredonia and the New York State Archives at Albany, New York, and converted into shapefile format, resulting in a total of 3,897 posts with 8,792 bearing trees for analysis by Wang (2007).

To characterize modern forest communities, we used data collected in the most recent complete FIA inventory of New York State, conducted between 1991 and 1993 (Alerich, Klevgard, and Miles 2004). FIA plots can be considered a fixed-area sample, although sampling scheme has varied somewhat historically and by state. No documentation regarding sampling scheme exists specific to our study area, but expansion factors contained within the data suggest that the area was inventoried using a combination of 0.5-ha fixed-area plots and variable-area sampling using a fifteen basal area factor prism. The FIA data contained twenty-one pairs and six triplets of apparently colocated plots. Within every such pair or triplet, each plot had the same coordinates but contained unique data with different numbers and species of trees. This suggested that they were separately located plots taken in the same general vicinity, rather than redundant data or return sampling of exactly the same location. Each such group was aggregated into a single plot for analysis purposes. This resulted in a total of 261 plots containing 4,303 individual trees.

One idiosyncrasy of the FIA data affecting spatial analysis is the intentional fuzzing of geographical coordinates by up to a mile and further swapping of up to 20 percent of plots on privately owned land within a county to protect plot integrity and landowner privacy (Alerich, Klevgard, and Miles 2004). *Fuzzing and swapping* means that fine-resolution analyses must be interpreted with caution. Comparison of PLSR and FIA data is thus often limited to coarse resolution analysis.

Taxa used by the HLC were associated with modern taxa following Wang, Kronenfeld, and Larsen (2009); some taxa were individual species, whereas others were groups of closely related species in the same genus. Although tree diameters are not recorded by the HLC township surveyors, estimates based on distances from survey post to bearing tree suggest that surveyors looked for trees larger than ∼9 inches in diameter at breast height (Kronenfeld and Wang 2007); therefore, this size cutoff was used for the FIA data as well. To avoid adverse effects of small sample size, only taxa that existed in both the HLC and the FIA surveys, and whose average abundance of the two time periods was ≥1 percent, were used in the analysis. A total of fourteen taxa met the criteria.

To derive gridded data sets for the HLC and FIA, relative abundances of each taxon that met the inclusion criteria were calculated at individual sample locations (HLC survey corners and FIA sample plots). The FIA data contain expansion factors signifying the number of trees per acre represented by each tree tallied, a number that varies according to the sampling scheme. To calculate relative abundances within an FIA plot, each tree was weighted by this expansion factor.

Grids of taxon abundance for each time period were created using kriging, a spatial interpolation method that has been used in previous presettlement vegetation reconstructions to allow visualization of tree taxon distributions in continuous representations (Brown 1998; Wang and Larsen 2006). Spatial interpolation also enables comparison of vegetation distribution derived from the HLC and FIA surveys that were different in numbers and locations of sample points. The spatially interpolated surfaces of taxa distribution were converted into 6 × 6 mile grid cells, a size corresponding to that of a township, which is the basic unit of the HLC survey. Substantial variation in taxon abundances within each of the 162 resultant grid cells might occur along topographic gradients, especially in the Appalachian Upland section; however, ecological modeling at this scale was not possible due to the inexact locations of FIA plot data. Conceptually, the interpolation process predicted taxon abundances as if the entire study region were forested. It did not capture the significant deforestation for agricultural and other land uses that occurs in parts of the study area, which should also be considered in conservation and management efforts.

Normalized average relative abundances of each taxon in the HLC and FIA gridded data are shown in Table 1. Decline in the abundance of beech and concomitant increases in red maple (*Acer rubrum*), poplar (*Populus spp.*), and black cherry (*Prunus serotina*) are the strongest components of change in taxonomic composition. Overall net change in taxonomic composition, measured as the sum of differences in overall taxon relative abundances divided by two was

Table 1 *Average relative abundances (in percent) of analyzed taxa in the Holland Land Company (HLC) and Forest Inventory and Analysis (FIA) gridded data sets*

Name used in text	Taxa	HLC	FIA
Ash	*Fraxinus sp.*	6.4	19.0
Basswood	*Tilia americana*	5.2	2.3
Beech	*Fagus grandifolia*	38.8	5.8
Birch	*Betula alleghaniensis*	2.6	1.3
Black cherry	*Prunus serotina*	0.4	10.4
Elm	*Ulmus americana*	4.5	3.5
Hemlock	*Tsuga canadensis*	9.0	5.6
Hickory	*Carya sp.*	1.3	3.8
Pine	*Pinus strobus*	2.6	1.7
Poplar	*Populus sp.*	0.6	10.9
Red maple	*Acer rubrum*	3.1	13.3
Red oak	*Quercus rubra*	0.2	4.3
Sugar maple	*Acer saccharum*	22.4	16.5
White oak	*Quercus alba*	3.0	1.6
Total		100	100

50 percent. This rate was similar to rates calculated from previous PLSR studies (e.g., Cowell 1998; Foster, Motzkin, and Slater 1998; Friedman and Reich 2005).

Classification, Map Production, and Analysis

Classification and statistical analysis of the gridded data were performed in a custom program written in Microsoft Visual Basic 6.0. Two conceptual models were created and compared: a *uniform model* in which each grid cell was considered to be occupied by a single forest type, and a *mixture model* in which multiple forest types were allowed to occupy each grid cell in various proportions. Classification under each model was represented numerically by a forest type definition matrix $\boldsymbol{C}$ and a proportion matrix $\boldsymbol{P}$, the latter constrained by Equations 3 and 4. In the uniform model, $\boldsymbol{P}$ was further constrained to exact values of zero or one.

Forest types under each model were defined using the k-means clustering algorithm, which seeks a predefined number (k) of data clusters that minimize within-group variance (MacQueen 1967). K-means clusters are sought from within the data itself, a conservative strategy for the mixture model given that many grid cells within our study area likely contained multiple forest communities. Using the same algorithm to derive forest type definitions for both models, however, allowed us to focus directly on the effects of within-cell mixing, rather than the more nebulous problem of defining community types.

The k-means clustering algorithm does not guarantee a unique outcome but depends somewhat on an initial set of arbitrarily determined seed clusters. To analyze algorithm variability and derive optimal community type definitions, 1,000 trial runs of the algorithm were conducted using randomly defined seed clusters for each of k= 2 to 8. This created a 2 × 7 × 1,000 matrix of algorithm runs on model (uniform vs. mixture), number of forest community types (k), and trial number. Input data for all algorithm runs consisted of the fourteen normalized relative taxon frequencies at 324 observations (162 grid cells from each of the HLC and FIA data sets).

For the uniform model, grid cells in each data set were assigned to the most similar forest type according to the inverse Euclidean distance metric of similarity. For the mixture model, proportions were assigned by linear unmixing.

To select forest type definitions from the 1,000 trials for each value of k, we calculated the *PVE* by each model as well as a measure of the degree of analog (d_A). *PVE* was calculated as:

$$PVE = 1 - \sum_{i=1}^{n}\sum_{j=1}^{m}(x_{ij} - \hat{x}_{ij})^2 \Big/ \sum_{j=1}^{m} var(x_{*j}) \tag{5}$$

where $var(x_{*j})$ denotes the variance of relative abundances of the jth taxon. For the uniform model, *PVE* is directly inversely related to the average squared Euclidean distance between each data point and its corresponding cluster, which is the criterion for the k-means algorithm (MacQueen 1967). Therefore, our methods maximize *PVE* for the uniform model but not the mixture model. *PVE* has a maximum value of one or 100 percent, indicating perfect correspondence between classification model and underlying data. Although there is no theoretical minimum, *PVE* = 0 is a logical bound that would result from assignment of every grid cell to a single forest type.

The degree of analog between presettlement and modern forest communities was expressed as an inverse function of the average net change in areal proportion of each forest type. Let $A_{q,t}$ denote the areal proportion of forest type q in time period t; the degree of analog (d_A) is defined as:

$$d_A = 1 - \sum_{q=1}^{k} \frac{|A_{q,HLC} - A_{q,FIA}|}{(A_{q,HLC} + A_{q,FIA})} \Big/ k \qquad (6)$$

Because net changes are weighted according to their magnitude relative to each forest type, this metric of global analog cannot be dominated by any single forest type. Note that d_A has a theoretical range of zero (no analog communities) to one or 100 percent (identical abundance of communities).

To enhance identification of analog communities under both the uniform and mixture models, the "optimal" solution from the trial runs of the k-means clustering algorithm for each value of $k = 2$ to 8 was defined as the trial that maximized the sum of $PVE + d_A$. Thus, we explicitly sought a classification that would result in maximum analog (combined with good model fit), but did so for both model types. The equal weighting of each metric is justified by the fact that they have similar theoretical ranges.

To further analyze differences in temporal continuity as portrayed by each model, we adopt the descriptive method of Pontius (2002) to calculate rates of *persistence*, *swap*, and *net change* for each forest type. These metrics provide a way to distinguish between changes in net quantity versus changes in spatial allocation (Pontius, Shusas, and McEachern 2004) and can be calculated from the composite fuzzy cross-tabulation matrix of Pontius and Cheuk (2006). This cross-tabulation matrix calculates the area of transition between each pair of forest types under the assumption that a forest type found in the same grid cell in both time periods occurs in the same location within the grid cell in each time period. Other assumptions are possible; Pontius and Connors (2009) describe methods for calculating a range of values for each cell in the cross-tabulation matrix. Of these, the composite cross-tabulation matrix results in maximal persistence (defined later).

Persistence describes the overall similarity in map pattern, and is defined as the percentage of grid cells that remain in the same forest type in both time periods. *Swap* indicates changes in spatial allocation but not quantity and is defined as the percentage of grid cells that transition to or from a given forest type, minus net gains and losses. The sum of persistence and swap thus indicates the component of each forest type that is globally present in both time periods. Finally, *net change* describes global fluctuations in the area occupied by each forest type. Net change is calculated as the sum of net gains, which equals the sum of net losses, and is equivalent to the rate of *turnover* described in many ecological studies (e.g., Cowell 1998). Together, these metrics provide a way to assess the spatial analog between presettlement and present.

We followed Dyer (2006) in using taxon indicator values to name forest types. Thematic maps were created for qualitative assessment and to aid in interpretation of the measures of persistence, swap, and net change. Previous work has shown that accurately interpreting color maps of mixture models requires a low degree of mixing (Kronenfeld 2005), which we did not encounter. To take advantage of the coarse spatial resolution of the gridded data, each forest type was depicted separately via fixed-angle line segments rotated around the center of each grid cell, with proportions indicated in grayscale. This gave a depiction of forest type proportions in both time periods, allowing visual analysis of the degree of overlap and disjunction.

Results

For classification schemes ranging from $k = 2$ to 8 forest types, degree of analog (d_A) was 8 percent higher, and *PVE* was 13 percent higher on average for the mixture model than for the uniform model for the trials that maximized the criterion ($PVE + d_A$). Figure 3 shows these values for $k = 3$ to 8. The results were consistent across all values of k except $k = 2$. This scheme was somewhat of an anomaly, as the trial selected by the optimization criterion had extremely low *PVE* for both models. This seemed to be an artifact of the k-means clustering algorithm, which produced two distinct groups of solutions for $k = 2$. This artifact did not appear for $k \geq 3$; the results of $k = 2$ are therefore omitted from further analysis.

The results varied substantially depending on the initial seeding of the k-means clustering algorithm. This is illustrated in Figure 4 for $k = 6$; other values of k resulted in a similar

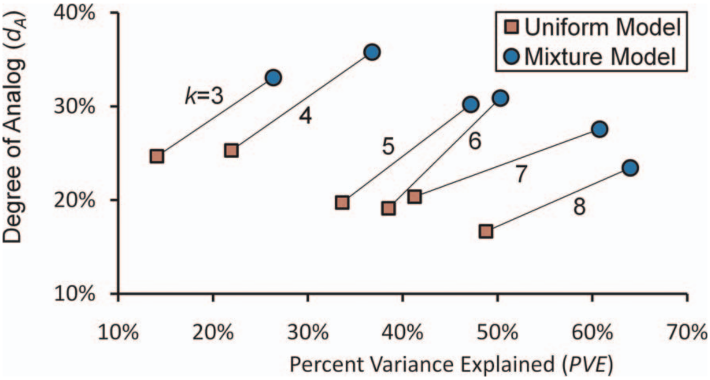

Figure 3 *Comparison of percentage variance explaned (PVE) and degree of analog (d_A) for uniform and mixture models. Values are for k-means clustering trial that maximized (PVE + d_A) for each of k = 3 to 8 forest types (number of forest types shown on line connecting each pair).*

pattern. Although a few trials of the uniform model resulted in higher *PVE* and d_A than a few trials under the mixture model, the overall pattern for the data clouds mirrors that of the optimally selected solutions (circled in Figure 4). Several trials resulted in zero analog under the uniform model, but no trial resulted in less than 15 percent analog under the mixture model.

Overall measures of persistence, swap, and net change for each classification scheme (Figure 5) revealed different spatial dynamics for the two models. A greater proportion of the study area persisted locally in the same forest type across both time periods under the mixture model (average = 15.9 percent) than the uniform model (average = 4.1 percent) for all values of $k > 2$. The swap rate, which indicates relocation of forest types, averaged slightly lower for the mixture model (14.3 percent) than for the uniform model (16.9 percent), but the sum of persistence and swap was still consistently higher under the mixture model. The lower rate of net change under the mixture model

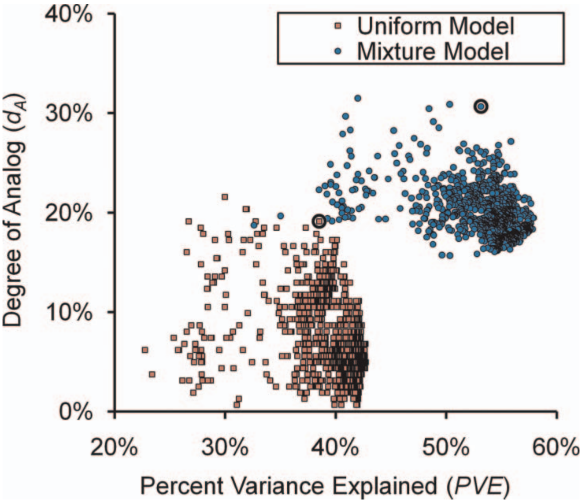

Figure 4 *Detailed view of explanatory power and degree of analog for 1,000 trials of k-means clustering algorithm for k = 6 using uniform and mixture models. Trials with highest value of optimization criterion (PVE + d_A) are circled.*

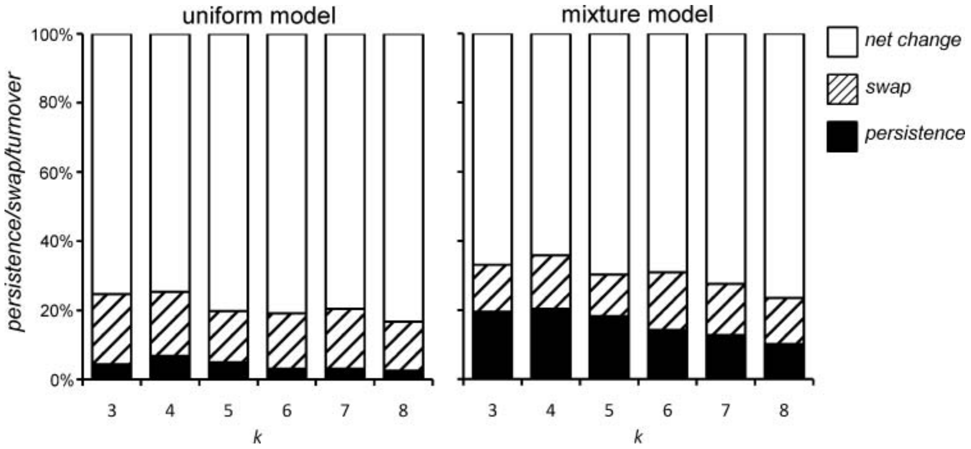

Figure 5 *Persistence, swap, and net change for optimally selected models for k = 3 to 8 forest types.*

in comparison to the uniform model therefore resulted entirely from increased persistence, as opposed to swap.

The optimally selected classification scheme for $k = 4$ is presented in further detail to illustrate the influence of the mixture model on individual forest community types. Both models led to similar forest type definitions (Table 2), with the same indicator taxa for three out of four forest types: (1) beech–sugar maple, (2) ash–poplar, and (3) sugar maple–black cherry. The fourth forest type also had similar taxon composition, but the indicator taxa for the uniform (red oak–red maple) and mixture (pine–white oak) models differed. For convenience, these two forest types will be referred to collectively as (4) oak–pine–red maple. Note that the taxa with the highest indicator values, used to name each forest type, were not always those with the highest relative abundances because indicator values also take into account the degree to which a taxon is found exclusively in one forest type and not other forest types.

Net loss, persistence, swap, and net gain for individual forest types are shown in Figure 6. Net loss or gain was lower, and persistence was higher, for every forest type under the mixture model than under the uniform model. Swap

Table 2 *Forest types defined in terms of typical relative taxon abundances (in percent) by the optimally selected four-cluster solutions (indicator taxa for each forest type are highlighted in bold)*

	Uniform model				**Mixture model**			
Taxa	**Beech–Sugar maple**	**Sugar maple–Black cherry**	**Ash–Poplar**	**Red oak–Red maple**	**Beech–Sugar maple**	**Sugar maple–Black cherry**	**Ash–Poplar**	**Pine–White oak**
Ash	5.5	10.0	**24.9**	7.3	5.3	9.7	**25.9**	8.0
Basswood	4.9	4.5	2.9	1.8	5.1	4.5	2.8	2.3
Beech	**44.9**	14.7	5.0	12.7	**47.1**	13.8	4.5	19.2
Birch	2.6	2.1	1.2	2.2	2.4	2.1	1.1	2.8
Black cherry	0.4	**8.7**	9.0	7.1	0.4	**9.3**	8.1	5.7
Elm	3.9	4.4	5.0	1.3	3.9	5.2	4.9	1.6
Hemlock	9.7	5.7	5.6	7.6	8.9	5.4	5.1	10.7
Hickory	1.0	1.7	4.8	2.5	0.7	1.9	4.9	2.6
Pine	2.5	0.9	1.1	6.4	1.4	0.8	1.3	**6.8**
Poplar	0.4	3.8	**13.0**	5.9	0.4	3.7	**13.3**	5.2
Red maple	2.5	7.2	12.1	**17.4**	1.8	7.9	12.8	12.6
Red oak	0.2	1.7	2.4	**9.5**	0.2	1.9	3.0	5.3
Sugar maple	**20.7**	**35.1**	12.6	13.7	**21.9**	**33.9**	11.8	13.3
White oak	1.9	1.4	1.4	7.5	1.6	1.5	1.6	**5.8**

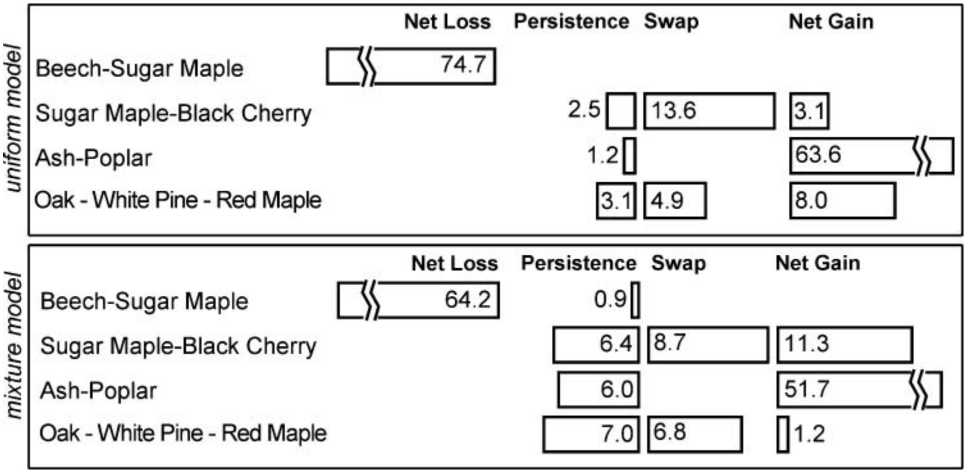

Figure 6 *Net loss, persistence, swap, and net gain (in percent) for optimally selected classification schemes with $k = 4$ forest types.*

occurred in only two forest types, sugar maple–black cherry and oak–pine–red maple. The effect of the mixture model on the rate of swap was inconsistent, increasing for sugar maple–black cherry but decreasing for oak–pine–red maple.

Figure 7 shows maps of both models, with proportions of each grid cell in a given forest type indicated by horizontal lines for the HLC (presettlement) and vertical lines for the FIA (modern) data sets. Persistence of a forest type within a grid cell is indicated by the presence of both horizontal and vertical lines, forming a cross. Such crosses are visually scarce in the uniform model (top row) but appear more frequently in the mixture model (bottom row). For example, ash–poplar, the most dominant forest type in the FIA, occupies only two (of a total 162) grid cells in the HLC under the uniform model but is present in thirty grid cells in the HLC under the mixture model. Sugar maple–black cherry presents a different pattern of change, but the influence of the mixture model is similar. Moderately abundant in both HLC and FIA, its center of distribution shifts to the south and west, persisting across both time periods in only four grid cells under the uniform model. Under the mixture model, however, the area of persistence expands to include portions of over thirty-five grid cells.

Space limitation precludes detailed analysis of other classifications. Although the specifics varied, the general tendencies observed for the case of $k = 4$ were similar across all values of k except $k = 2$, as described earlier.

Discussion

Assessments of whether or not communities have retained their coherence over time depend on the scale of analysis (Barnes et al. 1998). Researchers have suggested that the coarse data resolution of bearing trees requires PLSR data to be analyzed over at least several counties (Manies and Mladenoff 2000). Our results suggest that this coarse resolution, although unavoidable, creates a consistent bias against the discovery of analog communities unless nonuniformity within each spatial unit is accounted for. Net change in the areal extent of forest community types averaged 12 percent higher under the uniform model than when linear unmixing was used to estimate probable occurrence of communities below the resolution of analysis.

Even under the mixture model, net change in the areal extent of forest community types averaged 67 percent for the classification schemes we examined. This high *community* turnover rate is to be expected, considering the net change of 50 percent in overall taxon abundances. However, it also suggests that the changes in forest type composition are

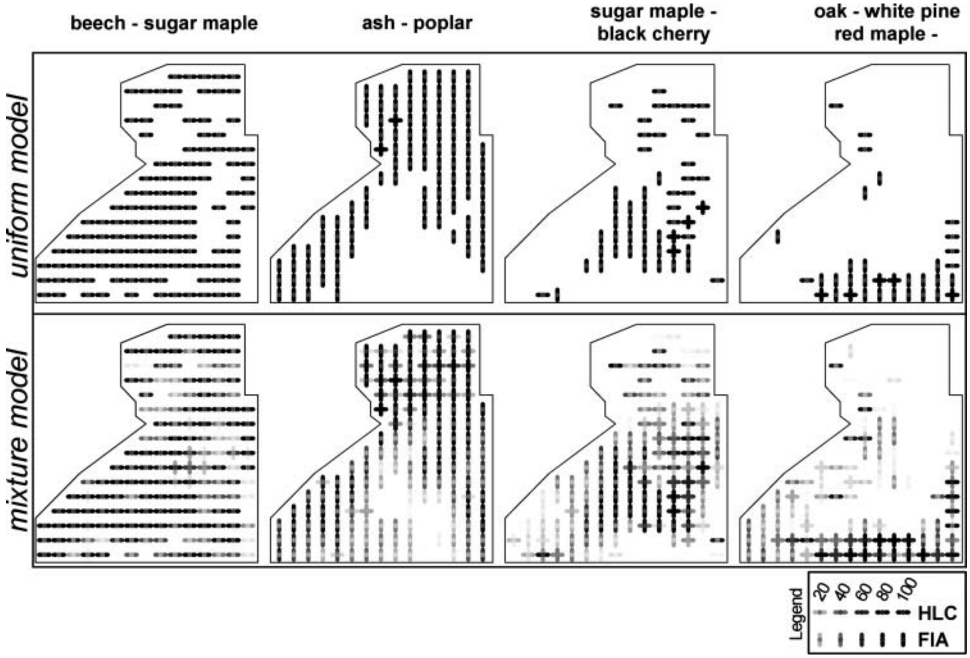

Figure 7 *Maps of forest types derived from uniform and mixture models. Shades of gray for horizontal and vertical bars indicate percentage of grid cell in given forest type for Holland Land Company (HLC) and Forest Inventory and Analysis (FIA data), respectively.*

primarily due to increases and decreases in individual taxon abundances, rather than changes in stand-scale associations between taxa. This might be true of other studies as well. For example, using a similar grid cell size as this study, Friedman and Reich (2005) suggested that close to two thirds of community types present in northeast Minnesota today have almost no presettlement analogs. Their data show a net change of 78 percent in the areal extent of forest community types, which is similar to the average of 79 percent we found under the uniform model across the classification schemes analyzed. Although their rate of net change in taxonomic composition of 34 percent is lower than ours, they also distinguished many more forest types, which would decrease the probability of finding analog communities in mixed data.

The difference of 12 percent in the net change in areal extent of forest types we observed under the uniform and mixture models had a disproportionate impact on the spatial continuity that could be observed for each forest type individually. Persistence of forest types within a given grid cell more than doubled under the mixture model in comparison to the uniform model (Figures 5 and 6). By revealing more detail, linear unmixing also provided a much clearer depiction of overall spatial dynamics, especially for the forest types that were less dominant within a given time period. For example, the mixture model presents an unambiguous picture of ash–poplar presence throughout the Lake Ontario Lowland province at the time of the HLC, only the scantest hint of which is visible under the uniform model (Figure 7).

One question that persists whenever linear unmixing is used is the degree of confidence that can be placed in the estimated subresolution proportions. Although the model fit, measured here by *PVE*, is a useful measure, there is no way to determine the precise composition of presettlement forest communities from coarse resolution data. Alternative definitions of forest types could be developed that would result in the same *PVE* as those presented here, as demonstrated by Wolbers and Stahel (2005). However, two arguments can be made that the general conclusions of this study are valid. First, because a conservative strategy (*k*-means clustering) was used to define forest types, the resulting forest type definitions were located within the centers of data clusters. Less conservative methods would have resulted in more extreme forest type definitions. For example, class definitions derived from the fuzzy *c*-means clustering algorithm (Bezdek, Ehrlich, and Full 1984) tend toward extreme values within the data matrix (McBratney and de Gruijter 1992), whereas the method of Lee and Seung (2001) results in forest type definitions outside of the data matrix entirely. More extreme forest type definitions would result in an even higher degree of mixing and thus a larger influence of the mixture model.

The observed spatial patterns of forest communities provide a second reason to be confident in the greater degree of continuity predicted by linear unmixing. The additional occurrences of each forest type "discovered" by linear unmixing in one time period showed a strong tendency toward locations in which the same forest type was dominant in the other time period. For example, the additional oak–pine–red maple predicted by linear unmixing to occur in the HLC data was concentrated along the Pennsylvania border, coinciding spatially with the area dominated by this forest type in the FIA data (Figure 7). This high degree of spatial correlation engenders confidence that linear unmixing captured meaningful information.

Although linear unmixing revealed greater continuity in forest communities in our study at the scale of two centuries across eight counties in western New York, this does not necessarily mean that the same results will be found in other study areas. Linear unmixing might in theory reveal either a greater or lesser degree of continuity than would otherwise be discerned. A useful area of future research would be to characterize the effects of mixture modeling on change analysis under different types of spatiotemporal pattern and, concomitantly, to characterize the range of underlying spatiotemporal patterns that could give rise to a given mixture model.

Conclusion

Previous studies have demonstrated large changes in tree taxon composition in the eastern United States since European settlement, which has led to a presumed lack of present-day analogs to presettlement forest communities. Both palynological and survey-based studies, however, are limited in resolution and therefore subject to the vagaries of generalization. Our findings suggest that the perceived lack of analog communities might result in part from data resolution, a problem that can be mitigated through the use of linear unmixing techniques. Although the increased analog component we found between presettlement and present was not large numerically, it substantially enhanced our ability to observe spatial patterns of change for individual forest community types. We recommend that unmixing techniques be used whenever community analysis is conducted using coarse resolution data. ■

Literature Cited

Abrams, M., and D. McCay. 1996. Vegetation-site relationships of witness trees (1780–1856) in the presettlement forests of eastern West Virginia. *Canadian Journal of Forest Research* 26:217–24.

Akerjord, M., and N. Christophersen. 1996. Assessing mixing models within a common framework. *Environmental Science and Technology* 30 (7): 2105–12.

Alerich, C., C. Klevgard, and P. Miles. 2004. *The Forest Inventory and Analysis database: Database description and users guide version 1.7.* Washington, DC: USDA Forest Service. http://ncrs2.fs.fed.us/4801/fiadb/ (last accessed 8 April 2007).

Bailey, R. 1995. *Description of the ecoregions of the United States.* 2nd ed. USDA Forest Service Miscellaneous Publication 1391. Washington, DC: USDA Forest Service.

Barnes, B., D. Zak, S. Denton, and S. Spurr. 1998. *Forest ecology.* 4th ed. New York: Wiley.

Batek, M., A. Rebertus, W. Schroeder, T. Haithcoat, E. Compas, and R. Guyette. 1999. Reconstruction of early nineteenth-century vegetation and fire regimes in the Missouri Ozarks. *Journal of Biogeography* 26:397–412.

Bezdek, J., R. Ehrlich, and W. Full. 1984. FCM: The fuzzy c-means clustering algorithm. *Computers and Geosciences* 10:191–203.

Black, B., H. Foster, and M. Abrams. 2002. Combining environmentally dependent and independent analyses of witness tree data in east-central Alabama. *Canadian Journal of Forest Research* 32 (11): 2060–75.

Bolliger, J., and D. Mladenoff. 2005. Quantifying spatial classification uncertainties of the historical Wisconsin landscape (USA). *Ecography* 28:141–56.

Bolliger, J., L. Schulte, S. Burrows, T. Sickley, and D. Mladenoff. 2004. Assessing ecological restoration potentials of Wisconsin (U.S.A.) using historical landscape reconstructions. *Restoration Ecology* 12 (1): 124–42.

Brown, D. 1998. Mapping historical forest types in Baraga County, Michigan, USA as fuzzy sets. *Plant Ecology* 134:97–111.

Burrough, P., and R. McDonnell. 1998. *Principles of geographical information systems.* New York: Oxford University Press.

Cogbill, C., J. Burk, and G. Motzkin. 2002. The forests of presettlement New England, USA: Spatial and compositional patterns based on town and proprietor surveys. *Journal of Biogeography* 29:1279–304.

Cowell, C. 1998. Historical change in vegetation and disturbance on the Georgia Piedmont. *American Midland Naturalist* 140:78–89.

Dyer, J. 2001. Using witness trees to assess forest change in southeastern Ohio. *Canadian Journal of Forest Research* 31 (10): 1708–18.

———. 2006. Revisiting the deciduous forests of eastern North America. *Bioscience* 56 (4): 341–52.

Fenneman, M. 1938. *Physiography of the eastern United States.* New York: McGraw-Hill.

Foody, G. M. 1996. Approaches for the production and evaluation of fuzzy land cover classification from remotely-sensed data. *International Journal of Remote Sensing* 17 (7): 1317–40.

Foster, D., G. Motzkin, and B. Slater. 1998. Land-use history as long-term broad-scale disturbance: Regional forest dynamics in central New England. *Ecosystems* 1 (1): 96–119.

Friedman, S., and P. Reich. 2005. Regional legacies of logging: Departure from presettlement forest conditions in northern Minnesota. *Ecological Applications* 15 (2): 726–44.

Fukuyama, Y., and M. Sugeno. 1989. A new method of choosing the number of clusters for the fuzzy *c*-means method. In *Proceedings of the Fifth Fuzzy Systems Symposium*, 247–50.

Gath, I., and A. Geva 1989. Unsupervised optimal fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11 (7): 773–81.

He, H., D. Mladenoff, T. Sickley, and G. Guntenspergen. 2000. GIS interpolations of witness tree records (1839–1866) for northern Wisconsin at multiple scales. *Journal of Biogeography* 27:1031–42.

Ichoku, C., and A. Karnieli. 1996. A review of mixture modeling techniques for sub-pixel land cover estimation. *Remote Sensing Reviews* 13:161–86.

Kronenfeld, B. 2005. Incorporating gradation as a communication device in area-class maps. *Cartography and Geographic Information Science* 32 (4): 231–41.

Kronenfeld, B., and Y. Wang. 2007. Accounting for surveyor inconsistency and bias in estimation of tree density from presettlement land survey records. *Canadian Journal of Forest Research* 37 (11): 2365–79.

Lee, D., and H. Seung. 2001. Algorithms for non-negative matrix factorization. *Advances in Neural Information Processing Systems* 13:556–62.

Leitner, L., C. Dunn, G. Guntenspergen, F. Stearns, and D. Sharpe. 1991. Effects of site, landscape features, and fire regime on vegetation patterns in presettlement southern Wisconsin. *Landscape Ecology* 5 (4): 203–17.

MacQueen, J. B. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability*, 281–97. Berkeley: University of California Press.

Manies, K., and D. Mladenoff. 2000. Testing methods to produce landscape-scale presettlement vegetation maps from the U.S. public land survey records. *Landscape Ecology* 15:741–54.

McBratney, A., and A. Moore. 1985. Application of fuzzy sets to climatic classification. *Agricultural and Forest Meteorology* 35:165–85.

McBratney, A. B., and J. J. de Gruijter. 1992. A continuum approach to soil classification by modified fuzzy *k*-means with extragrades. *Journal of Soil Science* 43:159–75.

Nascimento, S., B. B. Mirkin, and F. Moura-Pires. 2003. Modeling proportional membership in fuzzy clustering. *IEEE Transactions on Fuzzy Systems* 11 (2): 173–86.

Pontius, R. 2002. Statistical methods to partition effects of quantity and location during comparison of categorical maps at multiple resolutions. *Photogrammetric Engineering and Remote Sensing* 68 (10): 1041–49.

Pontius, R., and M. Cheuk. 2006. A generalized cross-tabulation matrix to compare soft-classified maps at multiple resolutions. *International Journal of Geographical Information Science* 20 (1): 1–30.

Pontius, R., and J. Connors. 2009. Range of categorical associations for comparison of maps with mixed pixels. *Photogrammetric Engineering and Remote Sensing* 75 (8): 963–69.

Pontius, R., E. Shusas, and M. McEachern. 2004. Detecting important categorical land changes while accounting for persistence. *Agriculture, Ecosystems and Environment* 101 (2–3): 251–68.

Power, C., A. Sims, and R. White. 2001. Hierarchical fuzzy pattern matching for the regional comparison of land use maps. *International Journal of Geographical Information Science* 15 (1): 77–100.

Radeloff, V., D. Mladenoff, H. He, and M. Boyce. 1999. Forest landscape change in the northwestern Wisconsin Pine Barrens from pre-European settlement to the present. *Canadian Journal of Forest Research* 29 (11): 1649–59.

Van Oort, P. A. J. 2005. Improving land cover change estimates by accounting for classification errors. *International Journal of Remote Sensing* 26 (14): 3009–24.

Wang, Y. 2005. Presettlement land survey records of vegetation: Geographic characteristics, quality and modes of analysis. *Progress in Physical Geography* 29:568–98.

———. 2007. Spatial patterns and vegetation-site relationships of the presettlement forests in western New York, USA. *Journal of Biogeography* 34:500–13.

Wang, Y., B. Kronenfeld, and C. Larsen. 2009. Spatial distribution of forest landscape change in western New York from presettlement to the present. *Canadian Journal of Forest Research* 39 (1): 76–88.

Wang, Y., and C. Larsen. 2006. Do coarse resolution U.S. presettlement land survey records adequately represent the spatial pattern of individual tree species? *Landscape Ecology* 21:1003–17.

Whitney, G., and J. DeCant. 2001. Government land office survey and other early land surveys. In *The historical ecology handbook: A restorationist's guide to reference ecosystems*, ed. D. Egan and E. Howell, 147–72. Covelo, CA: Island Press.

———. 2003. Physical and historical determinants of the pre- and post-settlement forests of northwestern Pennysylvania. *Canadian Journal of Forest Research* 33:1683–97.

Wolbers, M., and W. Stahel. 2005. Linear unmixing of multivariate observations: A structural model. *Journal of the American Statistical Association* 100 (472): 1328–42.

Wyckoff, W. 1988. *The developer's frontier: The making of the Western New York landscape*. London: Yale University Press.

Xie, X. L., and G. Beni. 1991. A validity measure for fuzzy clustering. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13 (8): 841–47.

Zhu, A. 1997. A similarity model for representing soil spatial information. *Geoderma* 77:217–42.

BARRY J. KRONENFELD is an assistant professor in the Department of Geography and Geoinformation Science at George Mason University, Fairfax, VA 22030. E-mail: bkronenf@gmu.edu. His research develops cartographic and statistical methods to analyze ecoregion patterns, presettlement land survey records, and point-based sampling schemes.

YI-CHEN WANG is an assistant professor in the Department of Geography, National University of Singapore, Singapore 117570. E-mail: geowyc@nus.edu.sg. Her research interests include reconstruction of historical forest landscapes from presettlement land survey records, patterns and processes of vegetation dynamics, and applications of geographic information systems and remote sensing in land use/cover change.

CHRIS P. S. LARSEN is an associate professor in the Department of Geography at University at Buffalo–SUNY, Buffalo, NY 14261. E-mail: larsen@buffalo.edu. His research uses a combination of modern, historical, and paleo data to explore the impacts of climate change, disturbance, and landscape configuration on the structure and composition of forests.