

Estimation of spatio-temporal correlations of prehistoric population and vegetation in North America

Bjoern Kriesche^a, Michelle A. Chaput^b, Rafal Kulik^c, Konrad Gajewski^b, Volker Schmidt^a

^a*Institute of Stochastics, Ulm University, Ulm, Germany*

^b*Laboratory for Paleoclimatology and Climatology, Department of Geography, Environment and Geomatics, University of Ottawa, Ottawa, Canada*

^c*Department of Mathematics and Statistics, University of Ottawa, Ottawa, Canada*

Email: bjoernkriesche@hotmail.de

Abstract: *We discuss a simple methodology to enable a statistical comparison of human population with the vegetation of North America over the past 13000 years. Nonparametric kernel methods are applied for temporal and spatial smoothing of point data obtained from the Neotoma Paleoecology Database and the Canadian Archaeological Radiocarbon Database, which results in sequences of maps showing the development of population and different plant taxa during the Holocene. The estimation of smooth spatial and spatio-temporal cross-correlation functions is proposed in order to detect relationships between population and vegetation in fixed time intervals. Furthermore, the effects of varying environment on demographic changes as well as potential impacts of populations on plant taxa over time are analyzed. Pointwise confidence bands for cross-correlation functions are computed and a robustness analysis is performed to assess the significance of obtained results. Considering the example of oak, an interpretation of our results for eastern North America shows the value of this methodology.*

Keywords: *Space-time analysis, Nonparametric estimation, Cross-correlation function*

1 INTRODUCTION

1.1 Motivation

In North America, European colonists had an enormous impact on the landscape since their arrival 400 years ago, for example, deforesting much of the landscape of Eastern North America by the early 20th century, although much has since regrown (Williams, 1989). Determining human impacts on the landscape of North America before historical times, however, is more complicated, and is done using (sub)fossil data. Two opposing viewpoints have been proposed about the nature of the vegetation of North America prior to the arrival of Europeans (e.g., Denevan, 1992; Vale, 2002). A first view is that North America in CE1492 was a “pristine landscape” with low population densities and vegetation largely unaltered by human activities. In this case, the primary factor causing changes in population numbers or in cultural/technological expression would have been climate and environmental changes. However, an alternate view suggests that in large parts of the Americas, First Nations used fire to clear the forest and maintain the prairies, practiced more extensive agriculture, and in general altered the forests and plains through extensive land use. Although local and regional studies have shown human impacts on the landscape (Delcourt and Delcourt, 2004; Munoz and Gajewski, 2010), it is important to understand this interaction at continental scales (e.g., Jäger and Neuhäusl, 1994). As this has not yet been discussed in the literature, we suggest to use simple and intuitive statistical strategies to examine this potential association over the course of the past 13000 years using centennial-scale time intervals.

1.2 Population intensity maps

We investigate space-time correlations of population density and vegetation abundances in prehistoric North America based on fossil data. While a methodology for the estimation of vegetation abundance is proposed in Sect. 4, we use population estimates that were derived in a previous study (Chaput et al., 2015). The original data come from the Canadian Archaeological Radiocarbon Database (CARD), which is a compilation of radiocarbon measurements that indicate the ages of samples from archaeological sites in North America (e.g., Gajewski et al., 2011). The CARD consists of 35,905 calibrated radiocarbon dates, with 29,609 of them containing suitable geographical, chronological and descriptive information linking them to culturally-distinct human activity. The suggested approach relies on the well-accepted assumption that the frequency of radiocarbon dates is, after accounting for sampling and taphonomic

biases, proportional to population activity (dates as data approach, e.g., Steele, 2010).

At first, a sequence of 121 500-year time intervals was selected, which range from 500-1000 BP (before present¹), 600-1100 BP, 700-1200 BP up to 12500-13000 BP. For each interval all corresponding radiocarbon dates are selected and spatially smoothed intensities of date counts are computed for all locations of the North American continent not permanently covered by ice using a nonparametric kernel density estimator with a two-dimensional Epanechnikov kernel and a globally fixed bandwidth of 600 km (the choice of this methodology is justified in Chaput et al. (2015)). In order to account for spatially inhomogeneous sampling strategies, we also estimate a sampling intensity based on all geographically distinct data locations in CARD (from all intervals) and divide intensities of date counts of each interval by sampling intensities. The resulting population intensities can be understood as (smoothed) numbers of dates per sampling site (falling in the considered time interval) and are interpreted as indicators of population density. In addition to sampling biases, the method also accounts for biases occurring due to taphonomic loss and boundary effects. Furthermore, some temporal smoothing is included in the method. Two examples of estimated population intensity maps are shown in Fig. 1. The maps show clear patterns of population activity and we consider that the results capture paleodemographic trends. This is justified by the observation that patterns correlate well with previous archaeological interpretations of population change across the North American continent during the Holocene that were derived partly from radiocarbon dates, but also other data sources and inferences.

1.3 Outline

In Sect. 2 we introduce the database used to obtain estimates of vegetation abundance of different taxa over the past 13000 years, including a discussion of calibrated radiocarbon ages. Sect. 3 describes an intuitive procedure for temporal smoothing and interpolation of pollen percentages to the target ages needed for the correlation analyses. The preparation of spatially smooth vegetation intensity maps is discussed in Sect. 4, along with a depiction of examples for one taxon and a comparison to existing results from the literature. In Sect. 5 we describe a simple approach to the estimation of spatial cross-correlations of vegetation abundances with population densities and the cross-correlations of changes in vegetation and populations at various temporal lags. This section also introduces a method to compute nonparametric confidence bands of estimated cross-correlations. In Sect. 6.1 we apply the presented

¹The term 'present' refers to the year CE1950, by convention. For example, the year 500 BP is equivalent to the year CE1450.

methods to a selected taxon and illustrate the insights obtained from this analysis. Furthermore, in Sect. 6.2, we perform a brief robustness study. Sect. 7 concludes.

2 THE NEOTOMA PALEOECOLOGY DATABASE

Space-time data of prehistoric pollen abundance used in this study are obtained from the Neotoma Paleocology Database (Grimm, 2008). Neotoma is a comprehensive compilation of fossil data from the Holocene, Pleistocene, and Pliocene for more than 8,400 sites worldwide.

2.1 Calibration of radiocarbon ages

For the estimation of vegetation intensities we use pollen data, which are typically acquired as follows. Consider a set of sites (lakes) on the North American continent, at which pollen data are available in Neotoma. Samples are taken along a sediment core for a sequence $d_0 < \dots < d_n$ of $n + 1$ depths. Each sample contains a certain number of fossil pollen, which are counted and classified. Next, the age of the sample at each depth needs to be determined. Since radiocarbon dating is rather expensive, it is not used to estimate the ages of all samples. Instead, radiocarbon ages are only determined for a subset $\{d_{i_1}, \dots, d_{i_j}\} \subset \{d_0, \dots, d_n\}$ and ages for depths $d \in \{d_0, \dots, d_n\} \setminus \{d_{i_1}, \dots, d_{i_j}\}$ are computed using interpolation methods². Ages obtained from the radiocarbon method (or from interpolation) need to be calibrated to be measurable in calendar years BP. As the manual use of standard calibration curves is not suitable for application to such a large number of dates as considered in our study, we suggest to convert radiocarbon ages into calibrated ages using a smoothed radiocarbon calibration curve (Grimm, 2008, Fig. 3). This simplified calibration is not exact but in Grimm (2008) it is estimated that the probability of the occurring error being less than or equal to 25 years is 0.47 and the probability of the occurring error being less than or equal to 200 years is 0.97. Since this is clearly below the temporal scale of this study, it is extremely unlikely that a significant bias is introduced by using the simplified calibration curve. The result of calibration is a sequence a_0, \dots, a_n of calibrated radiocarbon ages (the unit being cal years BP; for simplicity we will write BP) that correspond to the samples taken at depths d_0, \dots, d_n .

The correct procedure would be to calibrate radiocarbon ages at depths d_{i_1}, \dots, d_{i_j} first, resulting in calibrated ages a_{i_1}, \dots, a_{i_j} , and to interpolate ages for depths $d \in \{d_0, \dots, d_n\} \setminus \{d_{i_1}, \dots, d_{i_j}\}$ afterwards

²For more detailed information on age-depth modeling see, e.g., Bronk Ramsey (2008) or Haslett and Parnell (2008).

Correlations of population and vegetation in NA

based on a_{i_1}, \dots, a_{i_j} . For the majority of sites, however, ages are interpolated based on uncalibrated ages at d_{i_1}, \dots, d_{i_j} before feeding data into Neotoma, and we then calibrate the ages of all $n + 1$ depths d_0, \dots, d_n using the smoothed calibration curve. In general, this exchange of calibration and interpolation leads to an error in calibrated ages. Unfortunately, for those sites at which ages are interpolated before calibration, it does not seem possible to identify, which (uncalibrated) ages were obtained from dating and which from interpolation, making it impossible to eliminate this error. To investigate whether these calibration errors will significantly bias the results of our study, we perform the following comparison. We select 22 independent test samples from the literature, each consisting of a sequence d_0, \dots, d_n of depths in cm and a sequence $\tilde{a}_0, \dots, \tilde{a}_n$ of uncalibrated radiocarbon ages, with n varying between 4 and 17. For each test sample, we consider the sequence $\{\delta_1, \dots, \delta_k\} \subset [d_0, d_n]$, which contains all depths between d_0 and d_n that are a multiple of 5 cm. We first determine radiocarbon ages for depths $\delta_1, \dots, \delta_k$ by applying linear interpolation based on $\tilde{a}_0, \dots, \tilde{a}_n$ and afterwards calibrate interpolated ages using the smoothed calibration curve. Next, we first calibrate $\tilde{a}_0, \dots, \tilde{a}_n$ and then determine calibrated ages for $\delta_1, \dots, \delta_k$ by applying linear interpolation. This results in two calibrated radiocarbon ages a and a' for each depth $\delta \in \{\delta_1, \dots, \delta_k\}$, i.e., in two depth-age curves for each test sample. We find that the vast majority of absolute errors is smaller than 100 years with absolute errors of more than 300 years occurring extremely rarely (Fig. 2). Moreover, the largest errors occur for ages older than 13000 BP, which are not considered in our analysis. In summary, since the observed differences are small compared to the temporal scale of this study, we can assume that the errors occurring from exchanging calibration and interpolation of radiocarbon ages are negligible in the following.

2.2 Data selection

In order to access the Neotoma database for automatic data selection and processing we use the R package `neotoma` (Goring et al., 2015). First, we select all sites from Neotoma which are labeled with the geopolitical id 'Canada' or 'United States'. Then, all datasets associated with the sites obtained are loaded and those with the dataset type 'pollen' and the collection type 'composite' or 'core' are selected. Each dataset contains a sequence of $n + 1$ samples that are taken at depths $d_0 < \dots < d_n$ (resulting in n intervals between samples) and the corresponding ages a_0, \dots, a_n in cal years BP. The samples contain count data for different taxa. Since we are interested in pollen counts only, we select data with taxon group 'vascular plant', variable element 'pollen' or 'spore' and variable unit 'number of identified

specimen (NISP)'. Count data can sometimes not be compared directly across sites due to different levels of taxonomic resolution used by the researchers contributing to Neotoma. For example, one analyst might discriminate sub-taxa of *Acer* (maple), such as *Acer rubrum* (red maple) or *Acer saccharinum* (silver maple), while another might simply identify *Acer* to the genus level. To provide comparability, taxa and corresponding pollen data are aggregated using the standardization list suggested in Williams and Shuman (2008). Finally, the usual practice is to compute relative pollen abundances for all sites, ages and taxa based on pollen counts for all included plant taxa, which are much easier to interpret and compare. Altogether, we obtain a set of 1,151 sites, each of which contains a sequence a_0, \dots, a_n of calibrated radiocarbon ages together with the corresponding relative pollen abundances of 64 taxa.

3 TEMPORAL INTERPOLATION AND SMOOTHING OF POLLEN ABUNDANCES

In order to be able to estimate spatial vegetation intensity maps, relative pollen abundances need to be available at all 1,151 sites with Neotoma data simultaneously for the same years. Since relative pollen abundances of one taxon might vary considerably even during short time periods, some of which are random sampling errors, it is preferable to use smoothing methods over interpolation. In the following, we consider a fixed site, a fixed taxon and the corresponding relative pollen abundances. With $(a_0, p_0), \dots, (a_n, p_n)$, where $0 \leq a_0 < \dots < a_n$, we denote the ages of the available samples and the corresponding relative pollen abundances (taking values in $[0, 1]$) for the chosen site and taxon. We interpret $(a_0, p_0), \dots, (a_n, p_n)$ as (sorted) realizations of some independent and identically distributed random vectors³ $(A_0, P_0), \dots, (A_n, P_n)$ taking values in $[0, \infty) \times [0, 1]$. It seems almost impossible to find a parametric representation describing the dependence of random pollen abundances P_0, \dots, P_n on random ages A_0, \dots, A_n sufficiently well, which is why the following (nonlinear) relationship is assumed:

$$P_i = p(A_i) + \varepsilon_i \quad \text{for } i = 0, \dots, n, \quad (1)$$

where $p : [0, \infty) \rightarrow [0, 1]$ with $p(a) = \mathbb{E}(P_0 | A_0 = a)$ for $a \geq 0$ is the conditional expectation function of P_0 given A_0 and $\varepsilon_0, \dots, \varepsilon_n$ denote random errors with $\mathbb{E}\varepsilon_0 = \dots = \mathbb{E}\varepsilon_n = 0$. In order to

³In literature this setting is denoted as random design (e.g., Härdle et al., 2004). This framework seems suitable in our application since the ages of samples are not fixed a priori by the analyst performing the study and can thus be considered to be random. Furthermore, samples for different ages are acquired and investigated independently.

Correlations of population and vegetation in NA

estimate p by using kernel smoothing, we consider the time intervals I_1, \dots, I_n , where $I_i = [a_{i-1}, a_i]$ for $i = 1, \dots, n-1$ and $I_n = [a_{n-1}, a_n]$. We denote by $h_i = a_i - a_{i-1}$ the length of interval I_i for $i = 1, \dots, n$. The bandwidth h , which controls the degree of smoothing in kernel estimators, should be chosen not smaller than the maximum $\max\{h_1, \dots, h_n\}$. However, for some sites and taxa, long time periods I_i without a sample can occur leading to a bandwidth that can cause oversmoothing in other intervals I_j , $j \neq i$, eliminating too many details. To avoid such effects, we only consider those intervals $\{I_{i_1}, \dots, I_{i_k}\} \subset \{I_1, \dots, I_n\}$ that have a length of not more than 2,000 years and define $I = I_{i_1} \cup \dots \cup I_{i_k}$ and $h = \max\{h_{i_1}, \dots, h_{i_k}\}$. For all $a \in I$ an estimate $\hat{p}(a)$ of $p(a)$ can be determined using a (one-dimensional) Nadaraya-Watson estimator with the bandwidth h (e.g., Wand and Jones, 1995), by

$$\hat{p}(a) = \frac{\sum_{i=0}^n p_i K_1^G\left(\frac{a - a_i}{h}\right)}{\sum_{i=0}^n K_1^G\left(\frac{a - a_i}{h}\right)} \quad \text{for all } a \in I, \quad (2)$$

where $K_1^G : \mathbb{R} \rightarrow [0, \infty)$ denotes the one-dimensional Gaussian kernel defined as

$$K_1^G(u) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}u^2\right) \quad \text{for all } u \in \mathbb{R}. \quad (3)$$

Motivation for choosing the Gaussian kernel is that it has an unbounded support and is thus expected to provide a smooth estimate even in regions of sparse data. However, for $a \notin I$, using the Nadaraya-Watson estimator occasionally results in sudden decreases or increases making it an inappropriate choice. Therefore, we alternatively suggest to estimate p using linear interpolation, which leads to an estimate $\tilde{p}(a)$ of $p(a)$ for all $a \in [a_0, a_n]$. Finally, we set $\hat{p}(a) = \tilde{p}(a)$ for $a \in [a_0, a_n] \setminus I$. The proposed methodology is applied to 10 selected taxa. These include major taxa and those representative or characteristic of the major biomes (*Quercus* (oak): deciduous forest; *Picea* (spruce) and *Pinus* (pine): boreal forest; *Poaceae* (grass): Prairie), taxa used or potentially affected by human use (*Carya* (hickory), *Juglans* (walnut, butternut), *Castanea* (chestnut)), disturbance taxa (*Populus* (poplar, aspen)), or representatives of mature forests, i.e., non-disturbance (*Fagus* (beech), *Acer* (maple)). For example, if Native Americans extensively used fire to manage the forests, we would expect a positive association of human population densities and oak tree abundance (Abrams and Nowacki, 2008). If Native Americans managed the forests and planted useful trees, then there should be a spatial correlation of population density and nut trees,

Correlations of population and vegetation in NA

as seen on a local level (Wycoff, 1991). Munoz and Gajewski (2010) showed an increase in *Populus* in southern Ontario associated with introduction of agriculture, and we wanted to see if this occurred at a larger scale. On the other hand, we would also expect no association of spruce with human population densities, as human populations were assumed to be relatively low in boreal forests. In Fig. 3 sample data of relative pollen abundance from Neotoma are shown for two sites in North America together with estimates $\{\hat{p}(a), a \in I\}$ and $\{\tilde{p}(a), a \in [a_0, a_n]\}$ obtained using a Nadaraya-Watson estimator and linear interpolation, respectively. We find that choosing the smoothing parameter h as explained above leads to estimates that capture the temporal development very precisely (even for abrupt in- or decreases, see, e.g., the cyan curve in Fig. 3, top), whereas noise is also eliminated quite well, see, e.g., the yellow and the blue curve in Fig. 3 (top). In Fig. 3 (bottom) we have that $I \neq [a_0, a_n]$ since for ages between 3500 BP and 6500 BP interpolation is preferred over smoothing due to sparse data.

4 VEGETATION INTENSITY MAPS

4.1 Nonparametric estimation of vegetation intensity maps

We now address the question how spatially smooth vegetation maps can be determined. We consider the sequence of years 750 BP, 850 BP, \dots , 12750 BP, which corresponds to the midpoints of the 500-year intervals considered for the nonparametric estimation of population intensity maps sketched in Sect. 1.2. The estimation procedure is applied to each year and taxon separately. Let y be the considered year, i.e., $y \in \{750, 850, \dots, 12750\}$, and let W denote the set of all locations on the North American continent that are not covered by ice in year y . At first, we need to identify at which sites pollen abundances for the year y are available. In order to do that we consider the 1,151 sites with Neotoma data and check for each site individually whether $y \in [a_0, a_n]$, with a_0 and a_n being the site-specific minimal and maximal available calibrated radiocarbon ages introduced in Sect. 2.1. This results in a sequence $(s_1, \pi_1), \dots, (s_m, \pi_m)$, with $m \leq 1,151$, where s_1, \dots, s_m denote the sites in North America with available pollen abundances π_1, \dots, π_m in year y that were estimated using the kernel estimator in Eq. (2) or linear interpolation (for time intervals with sparse data). Typically, m is higher in more recent years since for older time periods there is a larger number of sites with $a_n < y$.

As relative pollen abundances may vary considerably among closely located sites (due to noise) and as we are mainly interested in large scale patterns of spatial vegetation intensities, it seems suitable to apply

Correlations of population and vegetation in NA

nonparametric methods for spatial smoothing as well. Since vegetation intensity maps are compared to population intensity maps in Sect. 5, we aim to provide as many similarities in the corresponding estimation procedures as possible. We suppose that $(s_1, \pi_1), \dots, (s_m, \pi_m)$ can be interpreted as realizations of some independent and identically distributed random vectors⁴ $(S_1, \Pi_1), \dots, (S_m, \Pi_m)$ with values in $W \times [0, 1]$. Again, it seems impossible to find a parametric representation that models the relationship between sites S_1, \dots, S_m and relative pollen abundances Π_1, \dots, Π_m sufficiently well, which is why we suppose that

$$\Pi_i = \pi(S_i) + \varepsilon'_i \quad \text{for } i = 1, \dots, m, \quad (4)$$

where $\pi : W \rightarrow [0, 1]$ with $\pi(s) = \mathbb{E}(\Pi_1 | S_1 = s)$ for $s \in W$ denotes the conditional expectation function of Π_1 given S_1 and $\varepsilon'_1, \dots, \varepsilon'_m$ are random errors with $\mathbb{E}\varepsilon'_1 = \dots = \mathbb{E}\varepsilon'_m = 0$. An extensive literature has demonstrated that pollen can be used as a quantitative index of past plant abundance (Birks and Birks, 1980), which is why the field $\{\pi(s), s \in W\}$ is considered as a map of expected intensities of the considered taxon in the following. However, as pollen percentages may be over- or underrepresented in comparison with the abundance of the plant on the landscape, we are not reconstructing vegetation, but rather mapping the change in time and space. We suggest to estimate $\{\pi(s), s \in W\}$ based on $(s_1, \pi_1), \dots, (s_m, \pi_m)$ using a two-dimensional Nadaraya-Watson estimator (e.g., Härdle et al., 2004). Accordingly, an estimate $\hat{\pi}(s)$ of $\pi(s)$ is determined by

$$\hat{\pi}(s) = \frac{\sum_{i=1}^m \pi_i K_2^E(s, s_i, h)}{\sum_{i=1}^m K_2^E(s, s_i, h)} \quad \text{for all } s \in W, \quad (5)$$

where $K_2^E : W \times W \times (0, \infty) \rightarrow [0, \infty)$ with

$$K_2^E(s_1, s_2, h) = \left(1 - \frac{d(s_1, s_2)^2}{h^2}\right) \mathbf{1}_{[0, h)}(d(s_1, s_2)) \quad \text{for all } s_1, s_2 \in W \quad (6)$$

⁴Choosing a random design is again suitable here. On the one hand, sampling sites were not designed by one analyst but chosen individually by many different researchers and can thus be considered to be independent random variables. On the other hand, corresponding pollen abundances are estimated independently for different sites based on the data from Neotoma, which are also sampled independently across sites.

Correlations of population and vegetation in NA

denotes a scaled two-dimensional Epanechnikov kernel, $h > 0$ is the bandwidth controlling the degree of smoothing and $d(s_1, s_2)$ denotes the geographic distance of two locations $s_1, s_2 \in W$ (Diggle and Ribeiro Jr., 2007, Sec. 2.7). We set $h = 600$ km in Eq. (5) to ensure comparability to the population intensity maps described in Sect. 1.2, which is also the reason for choosing the Epanechnikov kernel over a kernel with unbounded support (such as the Gaussian kernel). However, in contrast to the estimation of population intensity maps, we do not need to account for sampling biases and errors occurring due to taphonomic loss. On the one hand, the denominator in Eq. (5) prevents the estimator from being influenced by inhomogeneous sampling strategies. On the other hand, relative pollen abundances do not contain taphonomic biases as pollen does not degrade over the time-scale of our study. Computing $\{\hat{\pi}(s), s \in W\}$ for all $y \in \{750, \dots, 12750\}$ results in a sequence of estimated vegetation intensity maps. Note that elevation is currently not considered when estimating vegetation intensities as this would require a thorough extension of the methodology which will be developed in future studies within this research program.

4.2 Computation of taxon ranges

Most taxa analyzed in this paper have regions of typical occurrence (e.g., the eastern and southern U.S. for *Quercus*). Only in those regions do the estimated vegetation intensities have significantly positive values, whereas in the remaining areas intensities are zero or very close to zero (due to few pollen grains being transported by wind or caused by data and measurement errors). When comparing population and vegetation intensity maps in Sect. 5, it seems reasonable to only take into account the taxon range $\xi \subset W$. For example, if correlations between population activity and the intensity of *Quercus* are analyzed, only regions in the eastern and southern U.S. should be considered. In the western U.S. and in Canada (i.e., outside the range of *Quercus*), estimated vegetation intensities will be very close to zero although population activity is found in various regions, which will dilute correlation results. For that purpose we suggest an approach to determine (temporally varying) estimates of the taxon range ξ . Let $\{\hat{\pi}(s), s \in W\}$ be the estimated vegetation intensity map of a selected taxon for a given year y . Then, we suggest to compute an estimate $\hat{\xi}$ of the taxon range $\xi \subset W$ by

$$\hat{\xi} = \{s \in W : \hat{\pi}(s) \geq u \cdot \max\{\hat{\pi}(s), s \in W\}\}, \quad (7)$$

Correlations of population and vegetation in NA

with a suitable threshold $u \in (0, 1)$, i.e., the estimated range $\hat{\xi}$ consists of those locations on the North American continent, at which the local vegetation intensity is at least u times the global maximum of the vegetation intensity map. In order to determine an optimal choice for the threshold u , the taxon range is estimated for the most current year ($y = 750$ BP) for thresholds 0.05, 0.1, 0.15, 0.2, 0.25 and 0.3, and a visual comparison to the taxon's modern range is made (Thompson et al., 1999). We observe that for all 10 taxa considered in Sect. 3 the threshold $u = 0.2$ provides the best overall match. Since the estimate $\hat{\xi}$ as defined in Eq. (7) always depends on the maximum estimated vegetation intensity of the current year (and thus changes over time), we suppose that the suggested approach is also able to capture the typical taxon ranges for older years, which is confirmed by the examples shown in Sect. 4.3.

4.3 Interpretation of results

As an example illustrating the results of the methodology discussed here, we use the history of *Quercus* (oak) over the past 13000 years. Oak is the characteristic species of the Eastern Deciduous Forest and is a general indicator of the extent of this ecosystem. It is a major source of food for both humans and game animals (McShea and Healy, 2002). Prior to 12000 BP, high values of *Quercus* were restricted to Florida and the southern Gulf States. The range expanded and moved north and westward after 10750 BP, see Fig. 4 (top). At a smaller scale, the maps capture the lower values of *Quercus* in the upper slopes of the Appalachian Mountains, and *Quercus* became more abundant especially west of this mountain chain. During this time, *Quercus* remained abundant across all southern states. Beginning around 9000 BP, *Quercus* decreased in the south, and remained at higher values in a band across the Mid-Atlantic States, see Fig. 4 (center). During the period between 8000-7000 BP, maximum oak abundance remained in the region. Note that many sites have high values of oak pollen to the north of the region of maximum abundance, but other sites in proximity have lower values, so the map shows a decrease in regional abundance in this region. This pattern remained for the next few thousand years, followed by a slight decrease in abundance in the past 1500 years, see Fig. 4 (bottom).

Maps depicting the abundance of the pollen of tree taxa through time for eastern North America have been produced several times since the 1970s, as the database increased in size and methods to produce the maps have evolved, see the references in Williams et al. (2004). The latest version (Williams et al., 2004) used the North American Pollen Database with 759 sites, which has been incorporated and expanded in Neotoma. They made maps every 1000 to 2000 years using tri-cubic distance weighting to average

pollen data from a $300 \text{ km} \times 300 \text{ km} \times 500 \text{ m}$ (vertical) window to a 50 km grid. There is a very close visual correspondence of our maps to those of Williams et al. (2004); all of the features discussed above are seen in both sets of maps suggesting that the methodology proposed in the present paper is able to provide reliable indicators of vegetation abundance.

A more sophisticated approach to the inference of local vegetation intensities for tree taxa is discussed in Paciorek and McLachlan (2009), where Bayesian models and methods have been also used to estimate uncertainties in obtained results. However, we do not consider this to be necessary in our study as estimated vegetation intensity maps correspond particularly well with those of the existing literature, see above. Furthermore, the proposed methodology is closely related to that we used when estimating population intensity maps.

5 CORRELATION ANALYSIS OF POPULATION AND VEGETATION INTENSITIES

5.1 Cross-correlations of population and vegetation intensity maps

We aim to investigate whether significant correlations between vegetation and population intensities can be found. In the following, we again fix a taxon and a year $y \in \{750, 850, \dots, 12750\}$ and consider the corresponding map $\{\hat{\pi}(s), s \in \hat{\xi}\}$ of estimated vegetation intensities restricted to the estimated taxon range $\hat{\xi}$. Furthermore, let $\{\hat{\lambda}(s), s \in \hat{\xi}\}$ denote the corresponding estimated population intensity map for the 500-year time period $[y - 250, y + 250]$, which is obtained using a similar nonparametric smoothing method (Chaput et al., 2015). In statistical estimation theory, it is common to model estimators as random elements, which is why $\hat{\xi}$ is considered to be a realization of some random closed subset Ξ of W (Chiu et al., 2013). Furthermore, the estimates $\{\hat{\pi}(s), s \in \hat{\xi}\}$ and $\{\hat{\lambda}(s), s \in \hat{\xi}\}$ can be interpreted as realizations of two random fields $\Pi = \{\Pi(s), s \in W\}$ and $\Lambda = \{\Lambda(s), s \in W\}$ restricted to Ξ .

To infer probabilistic properties we assume that Π and Λ are jointly second-order stationary and isotropic (Wackernagel, 2003). This means that $\mathbb{E} \Pi(s) = \mu_{\Pi}$ and $\mathbb{E} \Lambda(s) = \mu_{\Lambda}$ for all $s \in W$, and that for any pair of locations $s_1, s_2 \in W$ the covariances $\text{cov}(\Pi(s_1), \Pi(s_2))$, $\text{cov}(\Lambda(s_1), \Lambda(s_2))$ and $\text{cov}(\Pi(s_1), \Lambda(s_2))$ only depend on the geographic distance $d(s_1, s_2)$ between s_1 and s_2 . Clearly such assumptions are hard to justify on a continental scale. However, when estimating cross-correlation functions we only consider the restriction of the random fields to the estimated taxon range, which covers a smaller region (in particular for *Quercus*) making this a more realistic assumption.

Correlations of population and vegetation in NA

For the analysis of relationships between vegetation and population intensities we consider cross-covariance and cross-correlation functions. Let $r_{max} > 0$ denote the maximum distance of any two locations in W , i.e., $r_{max} = \max\{d(s_1, s_2), s_1, s_2 \in W\}$. As Π and Λ are assumed to be jointly second-order stationary and isotropic, the cross-covariance function $C_{\Pi\Lambda} : [0, r_{max}] \rightarrow \mathbb{R}$ of Π and Λ is defined by $C_{\Pi\Lambda}(r) = \text{cov}(\Pi(s_1), \Lambda(s_2))$ for any $s_1, s_2 \in W$ with $r = d(s_1, s_2)$ (Genton and Kleiber, 2015). Since cross-covariance functions are difficult to interpret, they are typically normalized to obtain cross-correlation functions. Using that the variances $\sigma_{\Pi}^2 = \text{var} \Pi(s)$ and $\sigma_{\Lambda}^2 = \text{var} \Lambda(s)$ do not depend on $s \in W$ and by assuming that $\sigma_{\Pi}^2, \sigma_{\Lambda}^2 > 0$, the cross-correlation function $\rho_{\Pi\Lambda} : [0, r_{max}] \rightarrow [-1, 1]$ of Π and Λ can be defined by

$$\rho_{\Pi\Lambda}(r) = \frac{C_{\Pi\Lambda}(r)}{\sqrt{\sigma_{\Pi}^2 \sigma_{\Lambda}^2}} \quad \text{for all } r \in [0, r_{max}]. \quad (8)$$

In order to give an estimator of the cross-correlation function $\rho_{\Pi\Lambda}$, we first need to choose a finite sequence $t_1, \dots, t_k \in \hat{\xi}$ of sample points, at which the values $\hat{\pi}(t_1), \dots, \hat{\pi}(t_k)$ and $\hat{\lambda}(t_1), \dots, \hat{\lambda}(t_k)$ are determined to be used for estimation. For that purpose, we suggest to generate a realization of a homogeneous Poisson point process in W with intensity $\alpha = 0.0003$ (e.g., Chiu et al., 2013; Diggle, 2014), and use those points of the process as sample points t_1, \dots, t_k that fall into the considered estimated taxon range $\hat{\xi}$. The intensity $\alpha = 0.0003$ seems suitable as, on average, we get a number of sample points (3,876 in W in our example) that is high enough to ensure a reliable estimation but still allows computations to be done in a reasonable time.

The most intuitive approach is to first estimate the cross-covariance function $C_{\Pi\Lambda}$ using the method of moments (e.g., Genton and Kleiber, 2015). However, this method usually provides estimates for certain distance classes only leading to discontinuous and unstable estimated cross-covariance functions. Also the fitting of parametric models, as advised in most geostatistical applications to obtain smooth and continuous estimates (Montero et al., 2015), is not suitable here since it seems impossible to find a model that can be fitted adequately for all taxa and time periods. A more appropriate alternative is to consider a nonparametric kernel approach for the estimation of cross-covariance functions (similar to the Nadaraya-Watson estimator used in Sect. 3 and 4.1). This kind of estimator is frequently used in point process statistics for the estimation of mark correlation functions (Illian et al., 2008, Sect. 5.3), and has been

successfully applied and interpreted in a geographical context (e.g., Shimatani, 2002; Ledo et al., 2011)⁵.

Let $r_{est} > 0$ be chosen in such a way that for each $r \leq r_{est}$ the number of pairs of sampling points $t_i, t_j \in \{t_1, \dots, t_k\}$ in $\hat{\xi}$ with approximate distance $r \approx d(t_i, t_j)$ is reasonably large. Then, an estimate $\hat{C}_{\Pi\Lambda}(r)$ of $C_{\Pi\Lambda}(r)$ based on $\hat{\pi}(t_1), \dots, \hat{\pi}(t_k)$ and $\hat{\lambda}(t_1), \dots, \hat{\lambda}(t_k)$ can be computed by

$$\hat{C}_{\Pi\Lambda}(r) = \frac{\sum_{i=1}^k \sum_{j=1}^k (\hat{\pi}(t_i) - \hat{\mu}_{\Pi})(\hat{\lambda}(t_j) - \hat{\mu}_{\Lambda}) K_1^E \left(\frac{r-d(t_i, t_j)}{h} \right)}{\sum_{i=1}^k \sum_{j=1}^k K_1^E \left(\frac{r-d(t_i, t_j)}{h} \right)} \quad \text{for all } r \in [0, r_{est}], \quad (9)$$

where $\hat{\mu}_{\Pi} = \frac{1}{k} \sum_{i=1}^k \hat{\pi}(t_i)$ and $\hat{\mu}_{\Lambda} = \frac{1}{k} \sum_{i=1}^k \hat{\lambda}(t_i)$ are standard estimates of μ_{Π} and μ_{Λ} and $K_1^E : \mathbb{R} \rightarrow [0, \infty)$ denotes the one-dimensional Epanechnikov kernel with bandwidth $h > 0$, which is defined as

$$K_1^E(u) = \frac{3}{4}(1 - u^2)\mathbf{1}_{(-1,1)}(u) \quad \text{for all } u \in \mathbb{R}. \quad (10)$$

Comparisons of estimates based on different bandwidths have shown that $h = 20$ is a good choice to obtain smooth functions without eliminating important details. Using different types of kernel functions has a negligible effect on obtained results. Finally, a plug-in estimate $\hat{\rho}_{\Pi\Lambda}(r)$ of the cross-correlation function $\rho_{\Pi\Lambda}(r)$ is given by

$$\hat{\rho}_{\Pi\Lambda}(r) = \frac{\hat{C}_{\Pi\Lambda}(r)}{\sqrt{\hat{\sigma}_{\Pi}^2 \hat{\sigma}_{\Lambda}^2}} \quad \text{for all } r \in [0, r_{est}], \quad (11)$$

where $\hat{\sigma}_{\Pi}^2$ and $\hat{\sigma}_{\Lambda}^2$ are estimates of the variances σ_{Π}^2 and σ_{Λ}^2 . In order to obtain stable estimated cross-correlation functions it is recommended not to consider the standard moment estimates here. Instead, by using that $\sigma_{\Pi}^2 = \text{cov}(\Pi(s), \Pi(s))$ for all $s \in W$, an estimate $\hat{\sigma}_{\Pi}^2$ can be computed according to Eq. (9) for $r = 0$ with $\hat{\pi}(t_1), \dots, \hat{\pi}(t_k)$ and $\hat{\mu}_{\Pi}$ instead of $\hat{\lambda}(t_1), \dots, \hat{\lambda}(t_k)$ and $\hat{\mu}_{\Lambda}$. The estimate $\hat{\sigma}_{\Lambda}^2$ is determined analogously. The value $\hat{\rho}_{\Pi\Lambda}(r)$ for any $r \in [0, r_{est}]$ can be interpreted as the estimated correlation of the vegetation intensity and the population intensity at two arbitrary locations in the estimated taxon range $\hat{\xi}$ separated by distance r .

⁵A similar estimator for covariance functions has been proposed in Hall et al. (1994), where the authors consider an additional adjustment using the Fourier transform to ensure that estimated functions are positive semi-definite. We are not following this path, since the procedure is extremely computationally intensive and does not really affect qualitative interpretation.

5.2 Cross-correlations of population and vegetation changes with temporal lag

An even more interesting question that arises when analyzing correlations of vegetation and population is how certain taxa responded to changes in population and vice versa. For that purpose, we suggest to estimate and interpret cross-correlations of changes in vegetation and population intensity maps. We fix a taxon and a year $y \in \{1000, 1100, \dots, 12500\}$. Let $\hat{\pi}^{(y)}(s)$ denote the estimated change in vegetation intensity at location $s \in W$ between the years $y+250$ and $y-250$ (in years BP), which is computed based on the estimated vegetation intensities for the years $y+250$ and $y-250$. Analogously, let $\hat{\lambda}^{(y)}(s)$ be the estimated change of population intensity at s in the same period (or, to be more precise, between the 500-year intervals $[y, y+500]$ and $[y-500, y]$, whose midpoints correspond to $y+250$ and $y-250$). We only consider locations s that fall into the intersection $\hat{\xi}^{(y)}$ of the estimated taxon ranges corresponding to the years $y+250$ and $y-250$ in the following. By $\tilde{t}_1, \dots, \tilde{t}_{\tilde{k}}$ we denote those sample points (obtained from a realization of a Poisson point process, see Sect. 5.1) that fall into $\hat{\xi}^{(y)}$. We again suppose $\hat{\xi}^{(y)}$ to be a realization of some random closed subset $\Xi^{(y)}$ of W . Furthermore, we interpret the fields $\{\hat{\pi}^{(y)}(s), s \in \hat{\xi}^{(y)}\}$ and $\{\hat{\lambda}^{(y)}(s), s \in \hat{\xi}^{(y)}\}$ as realizations of some random fields $\Pi^{(y)} = \{\Pi^{(y)}(s), s \in W\}$ and $\Lambda^{(y)} = \{\Lambda^{(y)}(s), s \in W\}$ restricted to $\Xi^{(y)}$ with $\Pi^{(y)}(s), \Lambda^{(y)}(s)$ taking values in \mathbb{R} for $s \in W$.

It is conceivable that a change in vegetation intensity between the years $y+250$ and $y-250$ does not affect the change in population intensity in the same period but in a period a few hundred years later (or that the vegetation change is influenced by a population change some hundred years earlier). Therefore, we also analyze the cross-correlation function of $\Pi^{(y)}$ and $\Lambda^{(y+\delta)}$ for a temporal lag $\delta \in \{-1000, -900, \dots, 900, 1000\}$. At first, we again assume that the random fields $\Pi^{(y)}$ and $\Lambda^{(y+\delta)}$ are jointly second-order stationary and isotropic. Accordingly, the cross-covariance function $C_{\Pi^{(y)}\Lambda^{(y+\delta)}} : [0, r_{max}] \rightarrow \mathbb{R}$ of $\Pi^{(y)}$ and $\Lambda^{(y+\delta)}$ is defined by $C_{\Pi^{(y)}\Lambda^{(y+\delta)}}(r) = \text{cov}(\Pi^{(y)}(s_1), \Lambda^{(y+\delta)}(s_2))$ for $s_1, s_2 \in W$ with $r = d(s_1, s_2)$. The cross-correlation function $\rho_{\Pi^{(y)}\Lambda^{(y+\delta)}} : [0, r_{max}] \rightarrow [-1, 1]$ is defined as

$$\rho_{\Pi^{(y)}\Lambda^{(y+\delta)}}(r) = \frac{C_{\Pi^{(y)}\Lambda^{(y+\delta)}}(r)}{\sqrt{\sigma_{\Pi^{(y)}}^2 \sigma_{\Lambda^{(y+\delta)}}^2}} \quad \text{for all } r \in [0, r_{max}], \quad (12)$$

where $\sigma_{\Pi^{(y)}}^2 = \text{var} \Pi^{(y)}(s)$ and $\sigma_{\Lambda^{(y+\delta)}}^2 = \text{var} \Lambda^{(y+\delta)}(s)$ for any $s \in W$. For the estimation of $\rho_{\Pi^{(y)}\Lambda^{(y+\delta)}}$ consider a distance $r_{est}^{(y)} > 0$ such that for each $r \leq r_{est}^{(y)}$ the number of pairs of sampling points $\tilde{t}_i, \tilde{t}_j \in$

Correlations of population and vegetation in NA

$\{\tilde{t}_1, \dots, \tilde{t}_k\}$ in $\hat{\xi}^{(y)}$ with approximate distance $r \approx d(\tilde{t}_i, \tilde{t}_j)$ is large enough. For the computation of estimates $\hat{\rho}_{\Pi(y)\Lambda^{(y+\delta)}}(r)$ of the cross-correlations $\rho_{\Pi(y)\Lambda^{(y+\delta)}}(r)$ for all $r \in [0, r_{est}^{(y)}]$ we suggest to use the estimators proposed in Sect. 5.1 based on $\hat{\pi}^{(y)}(\tilde{t}_1), \dots, \hat{\pi}^{(y)}(\tilde{t}_k)$ and $\hat{\lambda}^{(y+\delta)}(\tilde{t}_1), \dots, \hat{\lambda}^{(y+\delta)}(\tilde{t}_k)$ instead of $\hat{\pi}(t_1), \dots, \hat{\pi}(t_k)$ and $\hat{\lambda}(t_1), \dots, \hat{\lambda}(t_k)$.

5.3 Nonparametric confidence bands of cross-correlation functions

In order to determine which of the values of estimated cross-correlation functions should be considered as significantly different from zero, we suggest to construct pointwise confidence bands using non-parametric resampling methods. We consider two approaches: a subsampling method and a bootstrap method (e.g., Chernick and LaBudde, 2011). We describe how the methods can be applied to estimated cross-correlations of vegetation and population intensity maps obtained in Sect. 5.1. An application to cross-correlations of vegetation and population changes, see Sect. 5.2, works analogously. Let $\{\hat{\rho}_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$ be the estimated cross-correlation function of some fixed year and taxon (with estimated range $\hat{\xi}$) obtained according to Eq. (11) based on estimated vegetation intensities $\hat{\pi}(t_1), \dots, \hat{\pi}(t_k)$ and estimated population intensities $\hat{\lambda}(t_1), \dots, \hat{\lambda}(t_k)$ at locations $t_1, \dots, t_k \in \hat{\xi}$. In the subsampling approach with subsampling proportion $\beta \in (0, 1)$ a random Bernoulli experiment with success probability β is performed for each sample point $t \in \{t_1, \dots, t_k\}$ (independently for different sample points) to decide whether the data pair $(\hat{\pi}(t), \hat{\lambda}(t))$ is retained or dropped. Based on all retained data pairs another estimate $\{\hat{\rho}_{\Pi\Lambda}^{(1)}(r), r \in [0, r_{est}]\}$ of $\{\rho_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$ is computed. When applying the bootstrap approach, $\{\hat{\rho}_{\Pi\Lambda}^{(1)}(r), r \in [0, r_{est}]\}$ is determined based on $\hat{\pi}(t'_1), \dots, \hat{\pi}(t'_k)$ and $\hat{\lambda}(t'_1), \dots, \hat{\lambda}(t'_k)$, where the k sample points t'_1, \dots, t'_k are drawn randomly with replacement from the set of original sample points $\{t_1, \dots, t_k\}$. The difference between both approaches can be interpreted as follows. In the subsampling approach, the cross-correlation function is re-estimated based on a reduced set of sampling points, i.e., we are interested in how the estimates change when it is assumed that some of the data are overrepresented or erroneous and are thus dropped. In the bootstrap approach, we do not only drop some of the sample points but also provide the estimated intensities at some other sample points with a higher weight to consider the case that those intensities might be underrepresented in the data. In both approaches the spatial correlation structure of the data is not violated since estimated vegetation and population intensities are still associated with the same geographical locations as before. The resampling procedure is repeated 5,000 times, resulting in a sample $\hat{\rho}_{\Pi\Lambda}^{(1)}(r), \dots, \hat{\rho}_{\Pi\Lambda}^{(5,000)}(r)$ of estimates for each $r \in [0, r_{est}]$. Then, based on this sam-

ple a confidence interval $[\underline{\theta}^{(\gamma)}(r), \bar{\theta}^{(\gamma)}(r)]$ of level $\gamma \in (0, 1)$ can be computed for $\hat{\rho}_{\Pi\Lambda}(r)$, where $\underline{\theta}^{(\gamma)}(r)$ denotes the empirical $(1 - \gamma)/2$ quantile and $\bar{\theta}^{(\gamma)}(r)$ the empirical $1 - (1 - \gamma)/2$ quantile of the sample $\hat{\rho}_{\Pi\Lambda}^{(1)}(r), \dots, \hat{\rho}_{\Pi\Lambda}^{(5,000)}(r)$. The confidence interval $[\underline{\theta}^{(\gamma)}(r), \bar{\theta}^{(\gamma)}(r)]$ is constructed in such a way that it contains $\gamma \cdot 100\%$ of the estimates $\hat{\rho}_{\Pi\Lambda}^{(1)}(r), \dots, \hat{\rho}_{\Pi\Lambda}^{(5,000)}(r)$. Finally, the functions $\{\underline{\theta}^{(\gamma)}(r), r \in [0, r_{est}]\}$ and $\{\bar{\theta}^{(\gamma)}(r), r \in [0, r_{est}]\}$ describe the (lower and upper) boundaries of pointwise confidence bands of level γ for the estimated cross-correlation function $\{\hat{\rho}_{\Pi\Lambda}(r), r \in [0, r_{est}]\}$. We consider an estimated cross-correlation $\hat{\rho}_{\Pi\Lambda}(r)$ to be significantly different from zero, if $0 \notin [\underline{\theta}^{(\gamma)}(r), \bar{\theta}^{(\gamma)}(r)]$, where usually $\gamma = 0.95$ or $\gamma = 0.99$ is chosen.

6 DISCUSSION OF RESULTS

6.1 Interpretation of correlation results

In this section we focus on the history of *Quercus* as a (more detailed) interpretation for all taxa considered in Sect. 3 would go beyond the scope of this paper. Fig. 5 shows estimated cross-correlation functions $\hat{\rho}_{\Pi\Lambda}$ for all years $y \in \{750, 850, \dots, 12750\}$, where warm colors indicate positive and cold colors negative cross-correlations. Prior to 10550 BP both *Quercus* and population intensities were very high in Florida and decreased to the north and westward, leading to high cross-correlations. During the period between 10550 BP and 6850 BP *Quercus* was moving northwards, while the population showed a complex pattern of changes. For example, relatively low values of *Quercus* in Texas while population intensities were high and decreasing values of population in Florida while *Quercus* remained stable contributed to low cross-correlations. This is a time period of quite some variability in the climate (Viau et al., 2006), although incompleteness of the population database cannot be ruled out. Between 6850 BP and 3650 BP, the Late Archaic cultural period, the cross-correlations were generally high. The distribution of *Quercus* did not change much during this time, and population gradually increased in the central portion of the continent, the region of maximum *Quercus* intensities. Fluctuations in the population caused the cross-correlations to increase and decrease during this time. In the past 3650 years, the Woodland Period, cross-correlations were uniformly high. Populations decreased in New England and greatly increased, within the central portion of the range, which continued to be the area of maximum *Quercus* abundances. The spatial association of the two variables in the late Holocene suggests optimal conditions for human population growth during this time.

Correlations of population and vegetation in NA

As explained in Sect. 5.2 we also estimate cross-correlation functions of changes in vegetation and population intensity during 500-year intervals. Additionally, a temporal lag is introduced, which implies that changes in *Quercus* in a certain interval are also compared to population changes in a interval starting and ending δ years earlier/later with $\delta \in \{-1000, -900, \dots, 900, 1000\}$. In order to allow for a thorough analysis of relationships between vegetation and population changes it is desirable to summarize cross-correlations for all temporal lags in one figure. For all $y \in \{1000, \dots, 12500\}$ and $\delta \in \{-1000, -900, \dots, 900, 1000\}$ we compute the mean of the estimated cross-correlations $\{\hat{\rho}_{\Pi(y)\Lambda(y+\delta)}(r), r = 30, 35, 40, \dots, 200 \text{ km}\}$ (to reflect local associations only) and summarize computed means in Fig. 6. Mean cross-correlations on the bold diagonal line correspond to the temporal lag $\delta = 0$, values above the diagonal line correspond to a negative temporal lag and values below the line to a positive lag. Several periods of positive cross-correlations between changes in intensities of *Quercus* and population are seen, notably between 11000 and 10000 BP, around 6000-5000 BP, and in the past 2000 years. Negative cross-correlations are found, e.g., between 7000 and 6000 BP.

The largest positive cross-correlations are found for changes in *Quercus* in the period between 11000 and 9500 BP and changes in population in 11300-10000 BP (for all positive and small negative temporal lags). This suggests that *Quercus* and population changed similarly over a period of more than 1000 years. During this time, both were decreasing in Florida, while increasing to the north (especially *Quercus*). This was a time of rapid warming in North America (Viau et al., 2006). During this time, people of the Paleoindian culture relied on a diverse set of resources including hunting different animals and gathering a variety of food (Fagan, 2000), and the increased diversity of the forest would have provided a suitable habitat to enable the increase of the population. However, the sample size used for estimating population intensities, especially to the south in Florida is small, so more definitive interpretation awaits further development of the database.

A period with low to medium negative associations between 8000-6500 was a time when both of the variables showed a complex series of changes. During this time, major changes in *Quercus* occurred in the west, due to a climate becoming moister, whereas changes in population occurred in various places, including to the east. In addition, a decrease in *Quercus* abundance in Florida and the southeast while population was increasing in the region contributed to the negative correlation. Further work is needed to ensure this is a real signal and not a function of lack of data; both datasets have low site densities

Correlations of population and vegetation in NA

in this region. The major changes in population at this time, therefore are most likely not driven by climate changes which would have been directly affecting *Quercus*. An increase in sedentism, and use of a diverse resource base by Middle Archaic peoples (Fagan, 2000) would have enabled adaptation to numerous environments, and contributed to a lower association with *Quercus* abundance. Delcourt and Delcourt (2004) have speculated that people may have thinned tree populations to encourage acorn yields, at least in some regions, and this would also contribute to a negative correlation.

In the recent past, agriculture and more sedentary lifestyles became increasingly established, with large cities in some areas and increasing impact on the environment (e.g., Fagan, 2000; Delcourt and Delcourt, 2004; Munoz and Gajewski, 2010). The population greatly increased across most of the area, especially in the past 2000 years and there were small increases in oak abundance, especially to the west. Agriculture and *Quercus* abundance would both be favored in areas with optimal environmental conditions, which could explain positive cross-correlations between change in *Quercus* in 3000-2500 BP and population in 2500-1500 BP as well as between change in *Quercus* in 1500-1000 BP and population in 2500-1500 BP. To conclude the interpretation and discussion we present some example results of confidence bands for cross-correlation functions as proposed in Sect. 5.3, where we focus on cross-correlations of changes in *Quercus* and population. Due to their time-consuming computation, confidence bands were only provided every 1000 years for temporal lag $\delta = 0$ using the bootstrap method and the subsampling method with success probabilities $\beta_1 = 0.25$, $\beta_2 = 0.5$ and $\beta_3 = 0.75$. In Fig. 7 confidence bands based on the subsampling method with $\beta_2 = 0.5$ for one cross-correlation function are shown. Cross-correlations for distances up to 550 km are significantly different from zero at significance level 0.99. Using the subsampling method with $\beta_1 = 0.25$ or $\beta_3 = 0.75$ results in wider or narrower bands, respectively. The bootstrap approach leads to bands that are almost identical to subsampling with $\beta_2 = 0.5$. In order to give a more general recommendation on how large (absolute) cross-correlations of changes in *Quercus* and population should be in order to be considered as significantly different from zero, we compute the maximum width of the confidence band (the difference of the upper and the lower bound of level 0.95) for all distances between 30 and 200 km (estimates for smaller distances tend to be occasionally unstable) for a sequence of estimated cross-correlation functions and divide them by 2 (Fig. 8). For example, the maximum of the green line indicates that when relying on the subsampling method with $\beta_2 = 0.5$ (which is a reasonably conservative choice) all cross-correlations greater than 0.35 (or smaller than -0.35) can be

assumed to be significant. Note that this approach assumes the confidence bands to be symmetric, which is the case (at least approximately) for the vast majority of estimated cross-correlation functions.

6.2 Robustness analysis

To conclude we perform several simulation studies to show the robustness of the obtained correlation results. In this context, we investigate to what extent the described signals are susceptible to uncertainties in the data or modifications of the methodology.

In Sect. 2.1 we mention two sources of error in the underlying pollen data which, however, are expected to have minor influence on correlation results. Altogether, there are five potential errors that can occur when compiling the data used in this paper: (1) the error from radiocarbon dating, (2) the error from radiocarbon calibration using, e.g., IntCal, (3) the error from using the simplified calibration curve from Grimm (2008) instead of IntCal, (4) the error from interpolation using a chronology, and (5) the error from exchanging interpolation and calibration. For each pollen sample considered in this study we simulate five independent random errors which are then added to the sample age. An (asymptotic) distribution of error (1) is estimated based on the raw ^{14}C errors in the CARD database, the distribution of error (2) is approximated based on the IntCal probability curves from a set of test samples, and error (4) is simulated based on the root mean squared errors of different chronologies considered in Telford et al. (2004). Errors (3) and (5) are simulated based on the histograms in Grimm (2008), Fig. 6 and Fig. 2 of this paper. While some of these distributions are rather rough approximations they still provide valuable insights on how different errors might impact the final results of radiocarbon dating. Based on this modified dataset Fig. 3-6 are recomputed. In all figures only minor differences are found, neither the temporal nor the spatial distributions of pollen abundances change noticeably. In particular, we observe that all correlation signals (including weaker ones) are retained in Fig. 5 and 6, which shows that the results described previously are strongly robust with respect to uncertainties in the pollen data.

In a second simulation study we investigate how modifications of the methodology for temporal and spatial smoothing affect correlation results. In this context we again recreate Fig. 3-6 where different kernel functions and bandwidths for temporal and spatial smoothing of the pollen data are used. At first, the temporal smoothing procedure described in Sect. 3 is repeated with modified bandwidths (50% or 200% of the original values) and using an Epanechnikov kernel, respectively. As expected reduced bandwidths

Correlations of population and vegetation in NA

lead to more fluctuations and do not eliminate noise sufficiently well while increased bandwidths smooth broadly but are not able to capture abrupt changes as for *Poaceae* in Fig. 3 (top). Thus, bandwidths as used in Sect. 3 are a good compromise. Changing the kernel function seems to have minor effect. Estimated vegetation intensity maps and cross-correlation functions, however, are only marginally affected by changing the parameters in temporal smoothing. All patterns described in Sect. 6.1 are retained, only some weaker correlation signals may vary slightly. In addition, we analyze the impact of modifying the spatial smoothing procedure described in Sect. 4.1. While using a two-dimensional Gaussian kernel has no noticeable effect (except of providing a smoother taxon range), changing the bandwidth can lead to stronger differences. The smaller the bandwidth the more fluctuations can be found and the rougher and smaller the taxon range, although most areas of strong population are still located in similar regions. Larger bandwidths lead to more extended taxon ranges and smoother patterns while the region of strongest population activity is moved slightly to the south. However, despite showing noticeable differences in population patterns, estimated cross-correlation functions are less strongly affected. In all scenarios we find the same correlation signals as described in Sect. 6.1, although they can be considerably weaker (small bandwidths down to 300 km) or slightly stronger (large bandwidths up to 1000 km). This shows clearly that the proposed methodology is robust to parameter configuration as well.

7 CONCLUSIONS

The aim of this study was to find out if there was a relation between human population intensities and plant abundance intensities over the past 13000 years in North America, using estimates of human population density from archaeological databases and plant abundance data from a pollen database. The methodology discussed in this paper is based on creating a comparable series of maps for the two datasets using simple nonparametric kernel approaches, and analyzing the spatio-temporal cross-correlations between them. Pointwise confidence bands of cross-correlation functions are proposed to investigate the significance of the obtained results.

Using an example of the relation between population densities in eastern North America and oak pollen abundance (an index of tree abundance through time), we could identify times in the past when there were both positive and negative associations between changes in the intensities of these two features. Inspection of the maps indicates times and locations where individual points seem to have undue influence on the correlations, especially in areas with few sites in the database. Future work refining the method could

Correlations of population and vegetation in NA

be to determine site density needed to more reliably estimate the intensities, especially to enable identification of overly influential or outlier points. Local analyses with smaller smoothing bandwidths may be effective here. In a companion study (Gajewski, Kriesche, Chaput, Kulik, and Schmidt, Gajewski et al.) the other taxa for which results have been prepared are analyzed in more detail. Since the community ecology of the tree species is well studied, and relative human use of the various taxa is also understood, consistencies between the taxa should help to identify the interpretable signal in these results. Finally these results will be interpreted in the context of more detailed regional studies to ensure that database limitations are not driving the correlations.

The relation between human population growth and ecosystem changes is complex, and causation can go in both directions. In addition, the human use of resources or human influence on the environment has changed through time as cultures evolve. This paper provides a method to quantify the relation between population and vegetation and show how they have changed through time and space. The approach is the first attempt to quantify this relation on continental scales and in that way contributes to our understanding of human-environment interactions in North America. Furthermore, the methodology has wide applications in geographic and environmental sciences, where large databases consisting of repeated measurements at a spatial system of sites are being accumulated.

REFERENCES

- Abrams, M. D. and G. J. Nowacki (2008). Native americans as active and passive promoters of mast and fruit trees in the eastern USA. *Holocene* 18, 1123–1137.
- Birks, H. J. B. and H. Birks (1980). *Quaternary Palaeoecology*. Caldwell: Blackburn Press.
- Bronk Ramsey, C. (2008). Deposition models for chronological records. *Quat. Sci. Rev.* 27(1-2), 42–60.
- Chaput, M. A., B. Kriesche, M. Betts, A. Martindale, R. Kulik, V. Schmidt, and K. Gajewski (2015). Spatiotemporal distribution of Holocene populations in North America. *Proc. Natl. Acad. Sci.* 112(39), 12127–12132.
- Chernick, M. R. and R. A. LaBudde (2011). *An Introduction to Bootstrap Methods with Applications to R*. Hoboken: J. Wiley & Sons.

Correlations of population and vegetation in NA

- Chiu, S. N., D. Stoyan, W. S. Kendall, and J. Mecke (2013). *Stochastic Geometry and its Applications* (3rd ed.). Chichester: J. Wiley & Sons.
- Delcourt, P. and H. Delcourt (2004). *Prehistoric Native Americans and Ecological Change*. Cambridge: Cambridge University Press.
- Denevan, W. (1992). The pristine myth: the landscape of the Americas in 1492. *Ann. Assoc. Am. Geogr.* 82, 369–385.
- Diggle, P. J. (2014). *Statistical Analysis of Spatial and Spatio-Temporal Point Patterns* (3rd ed.). Boca Raton: Chapman & Hall/CRC.
- Diggle, P. J. and P. J. Ribeiro Jr. (2007). *Model-Based Geostatistics*. New York: Springer.
- Fagan, B. M. (2000). *Ancient North America*. London: Thames and Hudson.
- Gajewski, K., B. Kriesche, M. A. Chaput, R. Kulik, and V. Schmidt. Human-vegetation interactions during the Holocene in North America. (submitted).
- Gajewski, K., S. Munoz, M. Peros, A. E. Viau, R. Morlan, and M. Betts (2011). The Canadian Archaeological Radiocarbon Database (CARD): archaeological ^{14}C dates in North America and their paleoenvironmental context. *Radiocarb.* 53, 371–394.
- Genton, M. G. and W. Kleiber (2015). Cross-covariance functions for multivariate geostatistics. *Stat. Sci.* 30(2), 147–163.
- Goring, S., A. Dawson, G. L. Simpson, K. Ram, R. W. Graham, E. C. Grimm, and J. W. Williams (2015). neotoma: a programmatic interface to the Neotoma paleoecological database. *Open Quat.* 1(2), 1–17.
- Grimm, E. C. (2008). Neotoma - an ecosystem database for the Pliocene, Pleistocene and Holocene. *Ill. State Mus. Sci. Pap. E Ser. 1*.
- Hall, P., N. I. Fisher, and B. Hoffmann (1994). On the nonparametric estimation of covariance functions. *Ann. Stat.* 22, 2115–2134.
- Härdle, W., M. Müller, S. Sperlich, and A. Werwatz (2004). *Nonparametric and Semiparametric Models*. Berlin: Springer.

Correlations of population and vegetation in NA

- Haslett, J. and A. Parnell (2008). A simple monotone process with application to radiocarbon-dated depth chronologies. *J. R. Stat. Soc. Ser. C* 57(4), 399–418.
- Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan (2008). *Statistical Analysis and Modelling of Spatial Point Patterns*. Chichester: J. Wiley & Sons.
- Jäger, K.-D. and R. Neuhäusl (1994). Interactions between natural environment and Neolithic man in Central Europe - an investigation based on comparative studies on vegetation and settlement with special emphasis on the view of natural science. In B. Frenzel (Ed.), *Evaluation of Land Surfaces Cleared from Forests in the Roman Iron Age and the Time of Migrating Germanic Tribes Based on Regional Pollen Diagrams*, pp. 75–81. Stuttgart: Gustav Fischer Verlag.
- Ledo, A., S. Condés, and F. Montes (2011). Intertype mark correlation function: a new tool for the analysis of species interactions. *Ecol. Model.* 222(3), 580–587.
- McShea, W. and W. Healy (2002). *Oak Forest Ecosystems*. Baltimore: John Hopkins University Press.
- Montero, J.-M., G. Fernández-Avilés, and J. Mateu (2015). *Spatial and Spatio-Temporal Geostatistical Modeling and Kriging*. Chichester: J. Wiley & Sons.
- Munoz, S. and K. Gajewski (2010). Distinguishing prehistoric human influence on late-Holocene forests in southern Ontario, Canada. *Holocene* 20, 967–981.
- Paciorek, C. J. and J. S. McLachlan (2009). Mapping ancient forests: Bayesian inference for spatio-temporal trends in forest composition using the fossil pollen proxy record. *J. Am. Stat. Assoc.* 104(486), 608–622.
- Shimatani, K. (2002). Point processes for fine-scale spatial genetics and molecular ecology. *Biom. J.* 44(3), 325–352.
- Steele, J. (2010). Radiocarbon dates as data: quantitative strategies for estimating colonization front speeds and event densities. *J. Archaeol. Sci.* 37(8), 2017–2030.
- Telford, R. J., E. Heegaard, and H. J. B. Birks (2004). All agedepth models are wrong: but how badly? *Quat. Sci. Rev.* 23, 1–5.

Correlations of population and vegetation in NA

- Thompson, R. S., K. H. Anderson, and P. J. Bartlein (1999). *Atlas of Relations between Climatic Parameters and Distributions of Important Trees and Shrubs in North America*. US Geological Survey Professional Paper 1650A-B. Denver: USGS.
- Vale, T. R. (2002). The pre-European landscape of the United States: pristine or humanized? In T. R. Vale (Ed.), *Fire, Native Peoples and the Natural Landscape*, pp. 1–40. Washington DC: Island Press.
- Viau, A. E., K. Gajewski, M. C. Sawada, and P. Fines (2006). Millennial-scale temperature variations in North America during the Holocene. *J. Geophys. Res.* 111(D09102).
- Wackernagel, H. (2003). *Multivariate Geostatistics* (3rd ed.). Berlin: Springer.
- Wand, M. P. and M. C. Jones (1995). *Kernel Smoothing*. London: Chapman & Hall.
- Williams, J. W. and B. N. Shuman (2008). Obtaining accurate and precise environmental reconstructions from the modern analog technique and North American surface pollen dataset. *Quat. Sci. Rev.* 27, 669–687.
- Williams, J. W., B. N. Shuman, T. Webb III, P. J. Bartlein, and P. L. Leduc (2004). Late-Quaternary vegetation dynamics in North America: scaling from taxa to biomes. *Ecol. Monogr.* 74, 309–324.
- Williams, M. (1989). *Americans and their Forests: A Historical Geography*. New York: Cambridge University Press.
- Wycoff, W. W. (1991). Black Walnut on Iroquoian landscapes. *Northeast Indian Quart.* 8, 4–17.

LIST OF FIGURES

- 1 Population intensity maps estimated from the CARD for two time intervals. Intensities are colored according to a logarithmic scale in the interval $[0, 1]$. Gray colors indicate areas permanently covered by ice.
- 2 Analysis of errors occurring when exchanging calibration and interpolation of radiocarbon ages. Left: histogram of errors for all chosen depths in all test samples. Right: depth-age-curves for three test samples showing the differences between correct procedure (blue) and exchanging calibration and interpolation (red).
- 3 Relative pollen abundances for two different sites in North America and 10 selected taxa: data from Neotoma (points), estimates obtained using linear interpolation (thin lines) and estimates obtained using a Nadaraya-Watson estimator (bold lines).
- 4 Estimated vegetation intensity maps showing the relative abundance of *Quercus* for three example years. The left-hand side illustrates intensities within the entire observation window W together with temporally smoothed abundances of *Quercus* at all sites from Neotoma with available data in the considered year. On the right only vegetation intensities within the taxon range estimated with threshold 0.2 are shown. Gray colors indicate areas permanently covered by ice.
- 5 Estimated cross-correlation functions for vegetation intensity maps of *Quercus* and population intensity maps for time intervals between 12750 BP and 750 BP.
- 6 Mean cross-correlations for 500-year changes in vegetation intensity maps of *Quercus* and population intensity maps. Means are computed based on cross-correlations for distances between 30 and 200 km. Values on the abscissa correspond to the midpoints of the 500-year intervals of change in population intensity and values on the ordinate correspond to the midpoints of the 500-year intervals of change in vegetation intensity.

Correlations of population and vegetation in NA

- 7 Estimated cross-correlation function of changes in *Quercus* and population between 9550 and 9050 BP together with pointwise confidence bands of levels 0.95 and 0.99 computed based on the subsampling method with success probability $\beta = 0.5$. The gray dashed line indicates the width of the confidence band (i.e., the difference of the upper and the lower bound) of level 0.95.
- 8 Half of the maximum width of confidence bands for cross-correlation functions of changes in *Quercus* and population. Bands are computed using four different approaches and maxima are determined among all widths for differences between 30 and 200 km. . .

Correlations of population and vegetation in NA

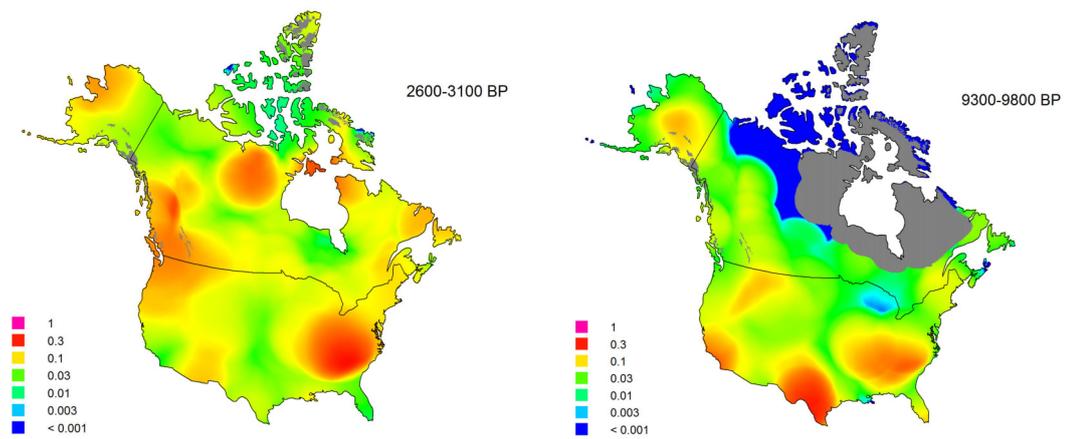


Figure 1. Population intensity maps estimated from the CARD for two time intervals. Intensities are colored according to a logarithmic scale in the interval $[0, 1]$. Gray colors indicate areas permanently covered by ice.

Correlations of population and vegetation in NA

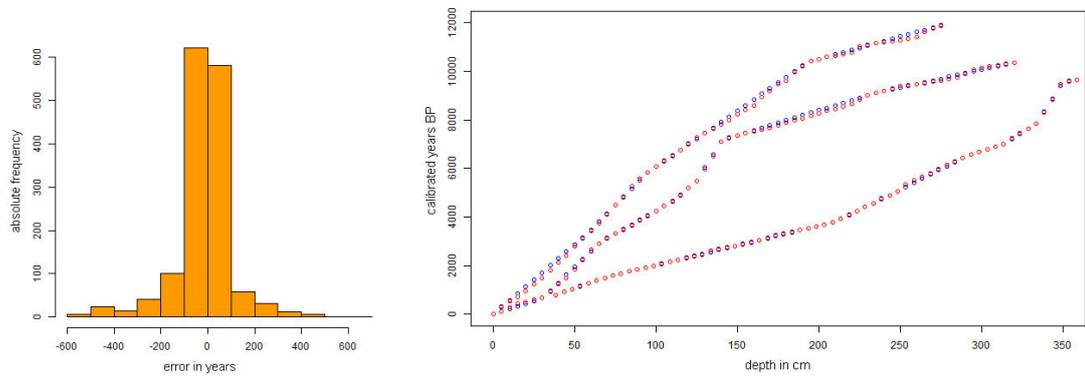


Figure 2. Analysis of errors occurring when exchanging calibration and interpolation of radiocarbon ages. Left: histogram of errors for all chosen depths in all test samples. Right: depth-age-curves for three test samples showing the differences between correct procedure (blue) and exchanging calibration and interpolation (red).

Correlations of population and vegetation in NA

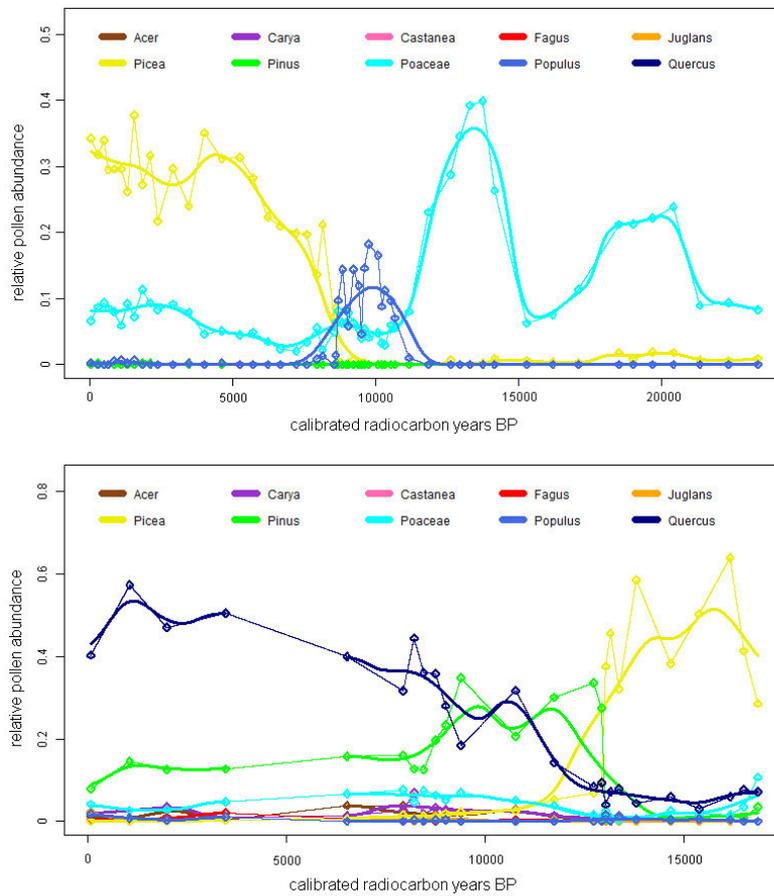


Figure 3. Relative pollen abundances for two different sites in North America and 10 selected taxa: data from Neotoma (points), estimates obtained using linear interpolation (thin lines) and estimates obtained using a Nadaraya-Watson estimator (bold lines).

Correlations of population and vegetation in NA

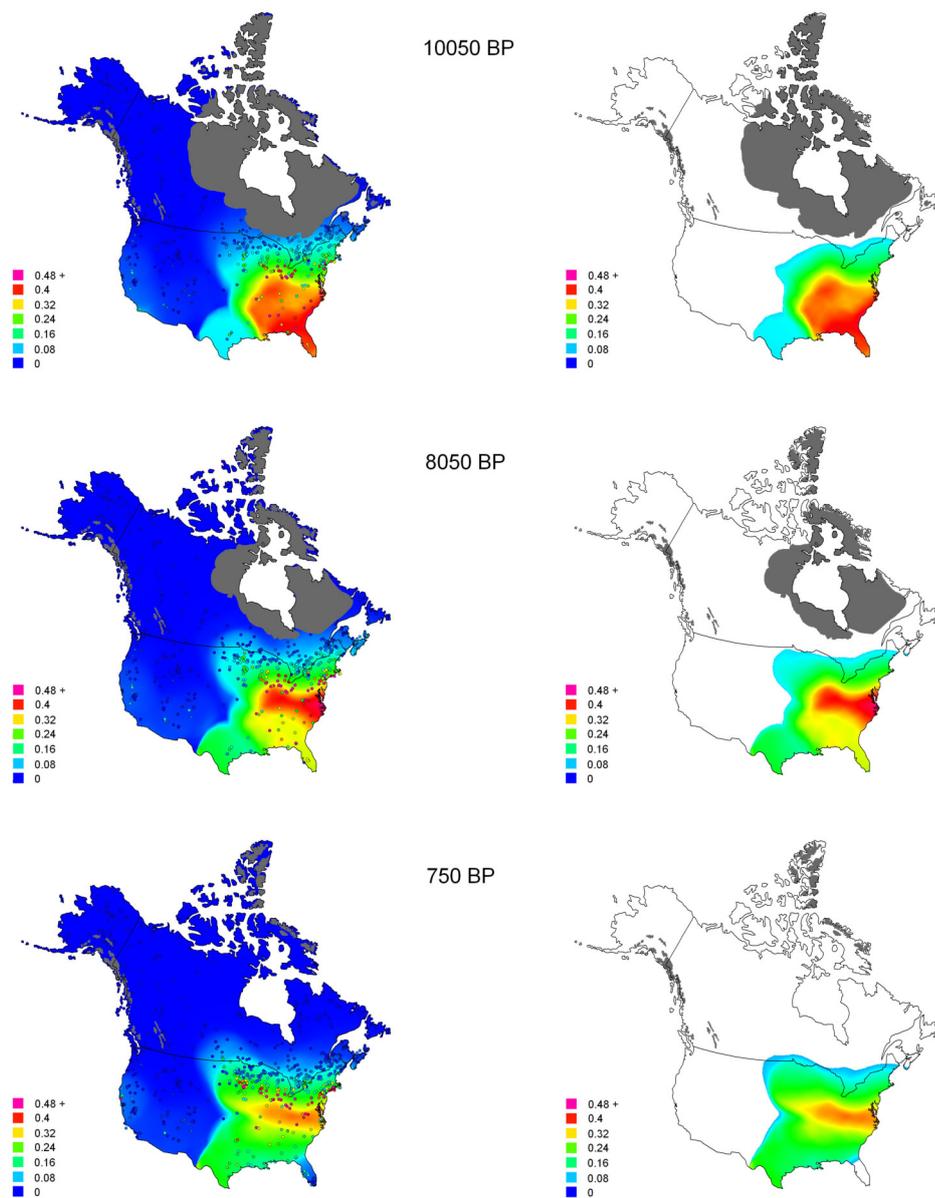


Figure 4. Estimated vegetation intensity maps showing the relative abundance of *Quercus* for three example years. The left-hand side illustrates intensities within the entire observation window W together with temporally smoothed abundances of *Quercus* at all sites from Neotoma with available data in the considered year. On the right only vegetation intensities within the taxon range estimated with threshold 0.2 are shown. Gray colors indicate areas permanently covered by ice.

Correlations of population and vegetation in NA

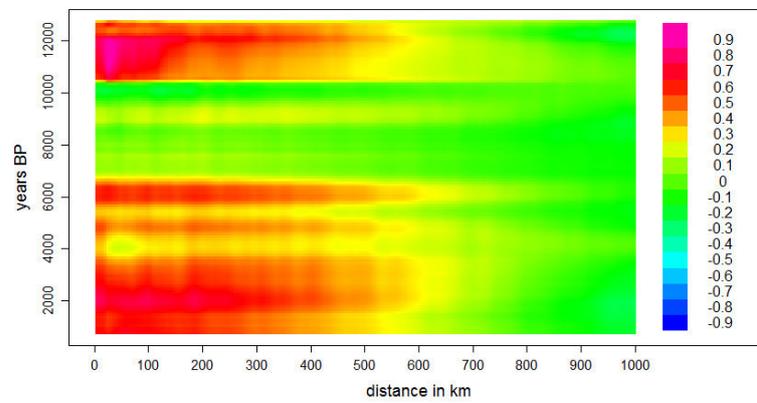


Figure 5. Estimated cross-correlation functions for vegetation intensity maps of *Quercus* and population intensity maps for time intervals between 12750 BP and 750 BP.

Correlations of population and vegetation in NA

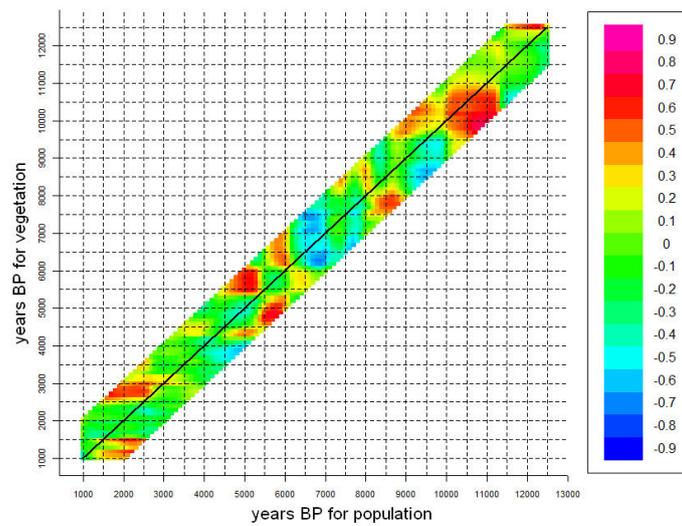


Figure 6. Mean cross-correlations for 500-year changes in vegetation intensity maps of *Quercus* and population intensity maps. Means are computed based on cross-correlations for distances between 30 and 200 km. Values on the abscissa correspond to the midpoints of the 500-year intervals of change in population intensity and values on the ordinate correspond to the midpoints of the 500-year intervals of change in vegetation intensity.

Correlations of population and vegetation in NA

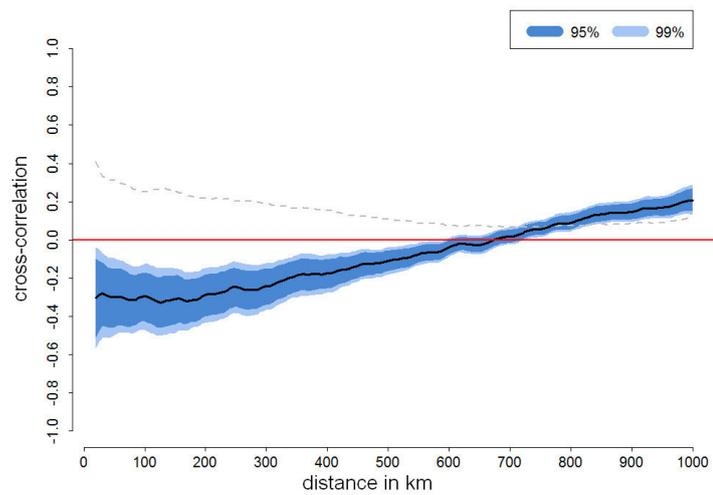


Figure 7. Estimated cross-correlation function of changes in *Quercus* and population between 9550 and 9050 BP together with pointwise confidence bands of levels 0.95 and 0.99 computed based on the subsampling method with success probability $\beta = 0.5$. The gray dashed line indicates the width of the confidence band (i.e., the difference of the upper and the lower bound) of level 0.95.

Correlations of population and vegetation in NA

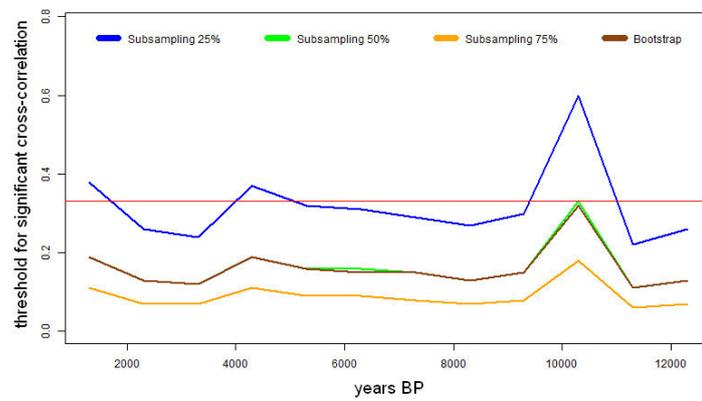


Figure 8. Half of the maximum width of confidence bands for cross-correlation functions of changes in *Quercus* and population. Bands are computed using four different approaches and maxima are determined among all widths for differences between 30 and 200 km.