Ecological Modelling 291 (2014) 109–120

Contents lists available at ScienceDirect

Ecological Modelling

journal homepage: www.elsevier.com/locate/ecolmodel

Using vegetation data within presettlement land survey records for species distribution modeling: A tale of two datasets

Stephen J. Tulowiecki *

Department of Geography, University at Buffalo, The State University of New York, Buffalo, NY 14261, USA

ARTICLE INFO

Article history:
Received 20 May 2014
Received in revised form 22 July 2014
Accepted 24 July 2014

Keywords:
Species distribution models
Presettlement land survey records
Bearing-tree data
Line-description data
Consensus modeling
Historical ecology

ABSTRACT

Researchers have long utilized the vegetation data within presettlement land survey records (PLSRs) to understand past forest composition in North America. PLSRs typically contain two datasets: bearing-tree (BT) data and line-description (LD) data. BT data are records of the trees that surveyors blazed adjacent to survey monuments, whereas LD data provide descriptions of tree species that surveyors observed along survey lines. Recently, studies have applied BT data to develop species distribution models (SDMs). SDMs create predictions of species distributions, based upon the modeled relationship between species presence and absence records, and environmental variables. Despite the applications of BT data in SDMs, the value of LD data for developing SDMs has not been explored. This study compares SDMs trained from LD data versus BT data, using PLSRs that were created ca. 1799–1814 CE in Chautauqua County, New York State. Using consensus modeling techniques, this study finds that despite positional uncertainty issues, LD data produce SDMs with better predictive performance than BT data, and more adequately generalize to independent datasets. Moreover, a comparable amount of data can be collected from LD data as from BT data, in order to develop models with greater predictive ability. This study challenges the use of BT data in SDMs, and suggests that modeling past species distributions can be accomplished more effectively using LD data.

1. Introduction

Ecologists and biogeographers have utilized presettlement land survey records (PLSRs) to understand past forests (Wang, 2005). PLSRs were constructed to demarcate townships and real property before European settlement predominately in North America (Schulte and Mladenoff, 2001). Many PLSRs contain two types of vegetation datasets, which are valuable in studying historical forest composition: line-description (LD) data and bearing-tree (BT) data (Fig. 1). In LD data, surveyors listed tree species that they observed along survey lines, oftentimes in order of abundance (Scull and Richardson, 2007). On the other hand, BT data is comprised of the point locations of individual trees that surveyors blazed and recorded, adjacent to survey monuments. Given their importance in reconstructing and understanding historical vegetation cover (Whitney, 1996), researchers have utilized both LD data and BT data in their studies, including to provide an ecological baseline under future climate change scenarios (Larsen et al., 2012).

To study the historical distribution of tree species, researchers have recently incorporated PLSR data into species distribution models (SDMs), aided by the use of geographical information systems (GIS). Arising from both statistical and machine learning methods (Breiman, 2001b; Franklin and Miller, 2009), SDMs identify relationships between recorded species records ("training data"), and environmental predictor variables. SDMs then use modeled relationships to predict the distribution of a species across space or time (Elith and Graham, 2009; Franklin and Miller, 2009). Numerous SDM algorithms have been developed and utilized, and the applications of SDMs have increased markedly in the last few decades (Elith et al., 2008; Peterson and Soberón, 2012).

Most studies that use PLSR data to train SDMs have utilized the point locations of BTs as training data, perhaps owing to its more spatially precise and less subjective characteristics (Wang, 2005). BT data have served as training data in numerous modeling approaches, such as hierarchical Bayesian models (He et al., 2007), weights-of-evidence (Fagin and Hoagland, 2011), boosted regression trees (Hanberry et al., 2012a,b), random forests (Hanberry et al., 2012c,d), and non-parametric multiplicative regression

Abbreviations: AUC, area under the receiver operating characteristic; BT, bearing-tree; HLC, Holland Land Company; LD, line-description; PLSR, presettlement land survey record; RLD, resampled line-description; SDM, species distribution model; TSS, true skill statistic.

* Tel.: +1 716 645 0471.
E-mail address: sjt7@buffalo.edu

http://dx.doi.org/10.1016/j.ecolmodel.2014.07.025
0304-3800/

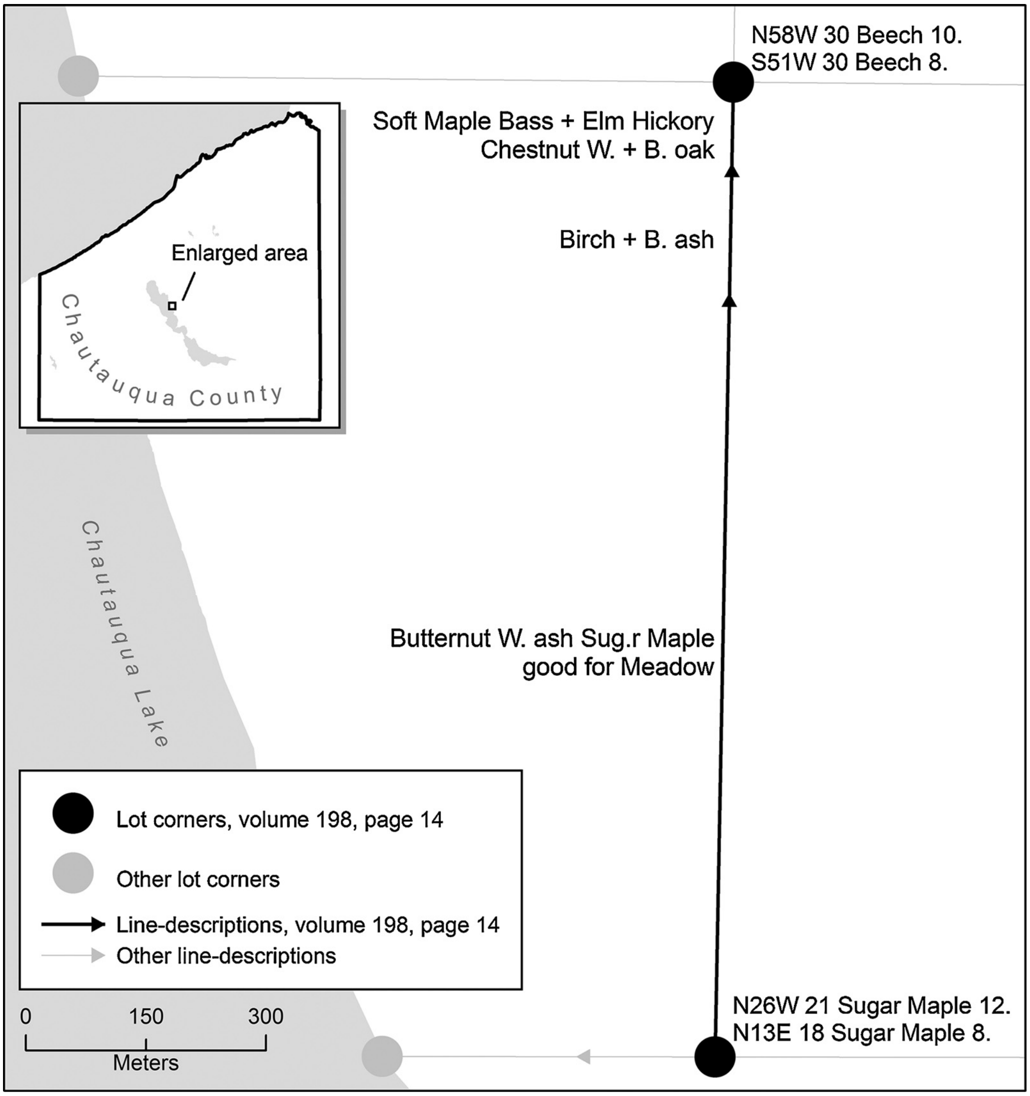

Fig. 1. An example of line-description (LD) data and bearing-tree (BT) data within PLSRs, transcribed from a single page of field notes in Chautauqua County. Surveyors recorded LDs of tree species that they observed along survey lines (e.g. "Birch + B. ash"). Surveyors also blazed and recorded BTs adjacent to posts. For example, "N26 W 21 Sugar Maple 12" indicates that a sugar maple (*Acer saccharum* Marsh.), 12 inches (30 cm) in diameter, was blazed and recorded at a distance of 21 links and bearing of N26 W from the post.

(Fahey et al., 2012). Using BT data, some studies have modeled presettlement vegetation at spatial grid cell resolutions as fine as 30 m (Yoo et al., 2013) to 120 m (He et al., 2007). Others have mapped historical forest cover, using geostatistical models of species presence and absence recorded in BT data (Manies and Mladenoff, 2000; Thomas-Van Gundy and Strager, 2012; Wang, 2007; Weih and Dick, 2008; Yoo and Trgovac, 2011).

Despite many applications of BT data, previous research has not explored the value of LD data for training SDMs, even though LD data contain many references to past vegetation (Fig. 1). In addition, no research has quantitatively compared SDMs trained from the two datasets, to examine which dataset yields SDMs of higher predictive performance. Puric-Mladenovic (2003) represents one example of using LD data to train SDMs, where the presence or absence of a species mentioned in a LD was designated at the centroid (or "midpoint") of the corresponding LD. The centroids that represented presence and absence of a species then served as point data to train SDMs. However, to the author's knowledge, this approach remains the only example of using LD data to train SDMs.

1.1. Line-description data and bearing-tree data: advantages and disadvantages

To compare the value of LD data and BT data for training SDMs, it is worthwhile to examine the spatial resolution, positional uncertainty, and bias issues of the two datasets. First, differences in spatial resolution and sampling exist between the two datasets. Surveyors often recorded LDs along survey lines at fine resolutions (Hutchison, 1988; Wang, 2005), and produced more references to species per unit area in LD data than in BT data (Seischab, 1992). LDs were sometimes aggregated in standardized lengths along survey lines (Batek et al., 1999; Scull and Richardson, 2007). In other PLSRs, such as Holland Land Company (HLC) surveys of Western New York, LDs typically corresponded with each landscape unit (e.g. a swamp) or soil quality (e.g. upland 1st quality) encountered by the surveyor (Wyckoff, 1988). In comparison to BT data, surveyors were permitted to list as many species as desired in LD data (Scull and Richardson, 2007). On the other hand, surveyors typically blazed and recorded BTs adjacent to survey monumentation at designated intervals (e.g. every 0.8 km) or at