Palaeogeography, Palaeoclimatology, Palaeoecology 442 (2016) 1–11

Contents lists available at ScienceDirect

Palaeogeography, Palaeoclimatology, Palaeoecology

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

Artificial neural networks reveal a high-resolution climatic signal in leaf physiognomy

Shu-Feng Li [a,e], Frédéric M.B. Jacques [a], Robert A. Spicer [c], Tao Su [a], Teresa E.V. Spicer [d], Jian Yang [d], Zhe-Kun Zhou [a,b,*]

[a] Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, China
[b] Key Laboratory of Biogeography and Biodiversity, Kunming Institute of Botany, Chinese Academy of Sciences, Kunming 650204, China
[c] Environment, Earth and Ecosystems, Centre for Earth, Planetary, Space and Astronomical Research, The Open University, United Kingdom
[d] State Key Laboratory of Systematic and Evolutionary Botany, Institute of Botany, the Chinese Academy of Sciences, Beijing 100093, China
[e] State Key Laboratory of Paleobiology and Stratigraphy, Nanjing Institute of Geology and Paleontology, Chinese Academy of Sciences, Nanjing 210008, China

A R T I C L E   I N F O

Article history:
Received 7 July 2015
Received in revised form 4 November 2015
Accepted 5 November 2015
Available online 10 November 2015

Keywords:
Artificial neural networks
Climate
CLAMP
CLANN
Fossil
Leaf physiognomy

A B S T R A C T

The relationship linking leaf physiognomy and climate has long been used in paleoclimatic reconstructions, but current models lose precision when worldwide data sets are considered because of the broader range of physiognomies that occur under the wider range of climate types represented. Our aim is to improve the predictive power of leaf physiognomy to yield climate signals, and here we explore the use of an algorithm based on the general regression neural network (GRNN), which we refer to as Climate Leaf Analysis with Neural Networks (CLANN). We then test our algorithm on Climate Leaf Analysis Multivariate Program (CLAMP) data sets and digital leaf physiognomy (DLP) data sets, and compare our results with those obtained from other computation methods. We explore the contribution of different physiognomic characters and test fossil sites from North America. The CLANN algorithm introduced here gives high predictive precision for all tested climatic parameters in both data sets. For the CLAMP data set neural network analysis improves the predictive capability as measured by R^2, to 0.86 for MAT on a worldwide basis, compared to 0.71 using the vector-based approach used in the standard analysis. Such a high resolution is attained due to the nonlinearity of the method, but at the cost of being susceptible to 'noise' in the calibration data. Tests show that the predictions are repeatable, and robust to information loss and applicable to fossil leaf data. The CLANN neural network algorithm used here confirms, and better resolves, the global leaf form–climate relationship, opening new approaches to paleoclimatic reconstruction and understanding the evolution of complex leaf function.

1. Introduction

With the increasing concern about global climate change, in recent decades there have been new and broader interests in paleoclimate reconstructions. Paleobotany has a long tradition of exploiting leaf form to determine past climates (e.g. Bailey and Sinnott, 1915, 1916; Dilcher, 1973; Greenwood and Wing, 1995; Jacobs, 1999, 2002; Jacques et al., 2011; Kowalski and Dilcher, 2003; Spicer and Herman, 2010; Srivastava et al., 2012; Su et al., 2013; Wilf, 1997; Wilf et al., 1998; Wing and Greenwood, 1993). These physiognomic methods have more than one hundred years of history from the first description of the relationship linking the percentage of leaves with entire margins to temperature (Britton and Brown, 1913). Since then both univariate (Wolfe, 1979) and multivariate approaches (Wolfe, 1990, 1993; Kovach and Spicer, 1996; Wolfe and Spicer, 1999; Spicer, 2000, 2007; Spicer et al., 2004, 2009; Jacques et al., 2011; Peppe et al., 2011; Yang et al., 2011, 2015) have been developed to reconstruct temperature, precipitation, and other climatic parameters.

There is a rich literature about the relationship between climate and foliar physiognomy: the percentage of species with entire margined leaves increases with temperature (Wolfe, 1979, 1993; Wilf, 1997), leaf size increases with moisture availability (Givnish, 1987; Peppe et al., 2011), and 'drip tips' are common in warm and humid environments (Leigh, 1975), but common mechanistic links between individual characters and single climate variables across all taxa remain elusive (Jordan, 2011). This is probably because modular genetic control, driven by pleiotropy, influences variation in form under a variable environment, and ultimately leads to natural selection for strongly linked but flexible functional systems (Falconer and Mackay, 1996; Juenger et al., 2005; Rodriguez et al., 2014) and "phenotypic integration," in which functionally related traits covary in complex ways within a given organism (Pigliucci, 2003). Leaves must optimize a variety of ecophysiological functions simultaneously and are developmentally integrated; it seems

* Corresponding author at: Key Laboratory of Tropical Forest Ecology, Xishuangbanna Tropical Botanical Garden, Chinese Academy of Sciences, Mengla 666303, China. Tel.: +86 691 8713226.
E-mail address: zhouzk@xtbg.ac.cn (Z.-K. Zhou).

http://dx.doi.org/10.1016/j.palaeo.2015.11.005
0031-0182/© 2015 Elsevier B.V. All rights reserved.

unlikely then that they would show single-character form–function–environment relationships (Yang et al., 2015).

A practical application of linking physiognomy and climate is the development of tools to retrodict past climate from leaf fossils via some form of function (f):

$$\text{Climate} = f(\text{physiognomic features})$$

The recent assembly of a large global foliar physiognomic data set (378 sites) demonstrates that in natural woody dicot vegetation, an integrated spectrum of leaf form exists across multiple leaf character states and species, and appears more strongly influenced by prevailing climate than biogeographic history (Yang et al., 2015). In this data set the co-variation of leaf traits across species suggests strong integration of leaf form (Yang et al., 2015). This work also demonstrates correlations between characters across a wide spectrum of woody dicot taxa despite the inclusion of samples from highly endemic floras. We know, therefore, that there is a relationship linking climate and physiognomy independent of taxonomic composition; however, we have little idea of the form of the function, how complex it is, and its parameters. Using simple relationships to build a complex multivariate function proves to be difficult because we lack information about how the factors interact. Univariate methods, such as leaf margin analysis, reduce the problem to one climatic parameter linked to one physiognomic feature, while digital leaf physiognomy (DLP) first looks at the physiognomic features with the highest explanation power, and then calculates the parameters of the function (Royer et al., 2005; Peppe et al., 2011). Both these approaches ignore, or in the case of DLP try to filter, the integrated nature of leaf form and function. If phenotypic integration results in an overall optimized solution to maximizing photosynthetic return for minimal resource investment, then the assumption that one particular subset of character/climate relationships is more important than another is dangerous when developing a climate proxy that has to be reliable across time and space.

CLAMP does not explicitly filter physiognomic characters but uses a vector-based direct ordination method, canonical correspondence analysis (ter Braak, 1986), to seek physiognomic/climate relationships across 31 leaf characters and a variety of climate variables. Like all previous approaches this uses traditional algebraic methods to compute model parameters. Major climate trends are sought through the cloud of modern natural or naturalized vegetation sites positioned relative to one another based on the leaf physiognomy displayed by at least twenty of their woody dicot component taxa. This cloud of calibration sites form what is known as 'physiognomic space'. By using observed climate data for each of the vegetation sites, climate trends across physiognomic space are determined and expressed as straight-line vectors. These vectors were originally aligned by eye in two-dimensional space (Wolfe, 1993) but subsequently objectively positioned first in two-dimensional space (Kovach and Spicer, 1996) and subsequently in four-dimensional space (Spicer et al., 2003). Higher dimensions carry little additional information for most calibration data sets. Samples with no known climate, such as fossil leaf assemblages, are positioned passively and their position along the vector (the vector score) is used to predict the unknown climate (CLAMP website: http://clamp.ibcas.ac.cn; Kovach and Spicer, 1996; Spicer, 2000; Wolfe and Spicer, 1999).

With small calibration data sets the structure of physiognomic space is relatively simple (Stranks and England, 1997; Spicer, 2000; Jacques et al., 2011), and the vector approach has proved adequate for predicting past climate accurately as measured against other paleoclimate proxies (Kennedy et al., 2002; Spicer et al., 2003), even accommodating some structural complexity by means of a non-linear regression model for calibrating the vectors. However, with large data sets spanning a diversity of vegetation and climates the ability of the vectors to capture the complexity of physiognomic space and the leaf form–climate relationship degrades (Yang et al., 2015), although the complexity can be visualized using a generalized additive model (Wood, 2011; Yang et al., 2015).

Because of the complexity of the relationship between plants and climate, it is quite likely that non-linear interactions exist among various aspects of the leaf physiognomy–climate relationship. So far, different approaches, such as CLAMP, DLP and other related modified approaches, seek linear trends that may constrain the prediction ability when worldwide data sets are considered. This is because a wider range of physiognomies occurs under the greater diversity of climate types represented as the size and geographic spread of the calibration data set increases. Non-linear relationships should be sought to improve the precision of paleoclimatic reconstruction from leaf physiognomy.

The purpose of this work is not to present an alternative paleoclimate proxy to those currently in use, but to explore a different way of revealing the information content of physiognomic space. In this study, we explore a new non-linear approach to approximate the function linking climate and physiognomy. The general regression neural network (GRNN) is a type of artificial neural network (ANN) that can approximate to both linear and nonlinear regressions (Specht, 1991). The GRNN is particularly advantageous with sparse data in a real-time environment, because the regression surface is instantly defined everywhere (Specht, 1991). As such the GRNN is a useful technique to investigate the climate and physiognomy relationship. We tested GRNN on two different physiognomy data sets and compared our results with those obtained from other computational methods. We also tested the GRNN using different physiognomic characters and fossil sites from North America.

2. Material and methods

2.1. Leaf physiognomy and climatic data sets

Two data sets were used in this study. The CLAMP global data set (Yang et al., 2015; the CLAMP website: http://clamp.ibcas.ac.cn) and the DLP data set (Peppe et al., 2011). Both data sets have a similar structure: a physiognomic data set that encapsulates leaf characteristics for each sampling site, and a meteorological data set describing the climate data for the same sites.

The CLAMP global data set used here is made up of 378 sites worldwide. The meteorological data usually consists of 11 parameters retrieved from a gridded data set (New et al., 2002; Spicer et al., 2009). The physiognomic data consist of a string of 31 characters describing leaf physiognomy across at least 20 taxa for each of those sites.

The DLP data set consists of 92 sites around the world. The meteorological data is made up of 10 parameters retrieved from WorldClim (Hijmans et al., 2005). The physiognomic data consist of 28 characters. Because the CLAMP data set is larger than the DLP data set, and thus potentially more complex, we chose to use the CLAMP data set in detailed tests of GRNN.

2.2. CLANN algorithm

We developed an algorithm based on GRNN. The predicted value (target) $\hat{Y}$ to input vector X in the GRNN is computed by the equation (Specht, 1991):

$$\hat{Y}(X) = \frac{\sum_{i=1}^{n} Y_i \, \exp\left(-\frac{D_i^2}{2\sigma^2}\right)}{\sum_{i=1}^{n} \exp\left(-\frac{D_i^2}{2\sigma^2}\right)} \quad (1)$$

where i is number of hidden nodes (samples) $i = 1, 2, 3, \ldots, n$. The optimal value of σ, which here denotes the spread, can be determined by cross-validation (Specht, 1991). D_i^2 is the Euclidian distance between