ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2020)

Global forensic geolocation with deep neural networks

Neal S. Grantham, Brian J. Reich, Eric B. Laber, Krishna Pacifici and Robert R. Dunn,

North Carolina State University, Raleigh, USA

Noah Fierer and Matthew Gebert

University of Colorado, Boulder, USA

and Julia S. Allwood and Seth A. Faith

North Carolina State University, Raleigh, USA

[Received July 2018. Revised May 2020]

Summary. An important problem in modern forensic analyses is identifying the provenance of materials at a crime scene, such as biological material on a piece of clothing. This procedure, which is known as geolocation, is conventionally guided by expert knowledge of the biological evidence and therefore tends to be application specific, labour intensive and often subjective. Purely data-driven methods have yet to be fully realized in this domain, because in part of the lack of a sufficiently rich source of data. However, high throughput sequencing technologies can identify tens of thousands of fungi and bacteria taxa by using DNA recovered from a single swab collected from nearly any object or surface. This microbial community, or microbiome, may be highly informative of the provenance of the sample, but data on the spatial variation of microbiomes are sparse and high dimensional and have a complex dependence structure that render them difficult to model with standard statistical tools. Deep learning algorithms have generated a tremendous amount of interest within the machine learning community for their predictive performance in high dimensional problems. We present DeepSpace: a new algorithm for geolocation that aggregates over an ensemble of deep neural network classifiers trained on randomly generated Voronoi partitions of a spatial domain. The DeepSpace algorithm makes remarkably good point predictions; for example, when applied to the microbiomes of over 1300 dust samples collected across continental USA, more than half of geolocation predictions produced by this model fall less than 100 km from their true origin, which is a 60% reduction in error from competing geolocation methods. Moreover, we apply DeepSpace to a novel data set of global dust samples collected from nearly 30 countries, finding that dust-associated fungi alone predict a sample's country of origin with nearly 90% accuracy.

Keywords: Citizen science; Machine learning; Microbiome; Non-homogeneous Poisson process; Spatial point pattern

1. Introduction

Forensic geolocation is the act of identifying the geographic source of an item on the basis of its observable properties. For example, soil on a suspect might be compared with soil at the location of a crime (e.g. Pye (2007)). Another promising source of evidence for use in geolocation is the

Address for correspondence: Brian J. Reich, Department of Statistics, North Carolina State University, 4264 SAS Hall, Box 8302, Raleigh, NC 27695, USA.
E-mail: bjreich@ncsu.edu

© 2020 Royal Statistical Society 0035–9254/20/69000

dust that accumulates on buildings, objects and individuals. The analysis of dust traces to aid forensic inquiry is not a new idea. In the 1920s, during his tenure as Director of the Laboratory of Police Technique in Lyon, France, Edmond Locard expressed excitement about the potential influence of dust-based forensic work, remarking that

> 'the microscopic debris that cover our clothes and bodies are the mute witnesses, sure and faithful, of all our movements and of all our encounters'

(Locard, 1930). Realization of Locard's vision, however, has been slow and episodic, particularly with regard to the biological materials in dust. A primary reason is that dust analysis has historically relied on the identification of microscopic grains of pollen or other biological materials (a laborious task) (Bryant and Jones, 2006). Such work is clearly useful, but inherently subjective and so dependent on the skill and perspective of the experts. Recent work has expanded the abilities of pollen for forensic analysis (Jones, 2012; Goodman *et al.*, 2015), but the technique remains more often used in archaeology than for forensics.

Dust can harbour deoxyribonucleic acid (DNA) from bacteria, archaea, fungi, plants and animals and we can leverage recent developments in high throughput DNA sequencing to assess the composition of these biotic assemblages (Barberán *et al.*, 2015b; Madden *et al.*, 2016; Craine *et al.*, 2017). Such DNA analyses can generate a more comprehensive biogeographical fingerprint than is possible with traditional morphology-based assessments of pollen alone. Advances in sequencing technology enable high fidelity identification of hundreds of thousands of taxa within samples containing biological material. Grantham *et al.* (2015) developed a preliminary forensic geolocation model for fungal occupancy (presence or absence) data built on Bayesian discriminant analysis (BDA) that compares a sample with a reference database and predicts the sample's provenance. When applied to data derived from dust samples from nearly 1000 homes across continental USA, the geolocation algorithm makes point predictions that fall within 230 km, on average, of their true provenance. These were the first results to estimate formally the extent to which dust-associated fungi can predict the provenance of a sample.

Although the initial results based on BDA are encouraging, the approach has limitations that may impede its geolocation performance. First, it assumes that microbial taxa in the microbiome occur independently of one another. However, taxa may have complex interconnectedness due to competition for resources or shared environmental preferences. Also, BDA uses all taxa, but not all taxa are informative about the sample's provenance and the noise of uninformative taxa obscures the geographic signal. Finally, the model does not make use of the well-established body of literature on spatial point patterns which could markedly aid model testing and future model development (Baddeley *et al.*, 2015).

In the spatial point pattern literature, points in space are often framed as arising from a non-homogeneous Poisson process with an unknown intensity surface distributed over the spatial domain (e.g. Baddeley *et al.* (2015) and Gelfand *et al.* (2010)). With respect to geolocation, interest lies in understanding how this intensity surface varies over space in terms of the covariate data. Liang *et al.* (2008) modelled the intensity as a log-Gaussian process and approximated it with a Monte Carlo algorithm using a knot-based predictive process. However, this approach is computationally infeasible when the covariate data are high dimensional. To handle spatial point patterns with high dimensional covariate data, lasso penalties for Poisson process data have been developed (Yue and Loh, 2015), but this requires covariates that are known spatial functions (unlike a sample microbiome) and assumes a log-linear effect.

In this paper, we propose to use deep learning to leverage microbiome data for geolocation. Deep learning has recently been applied successfully to related high dimensional problems. For example, deep neural networks have greatly improved human speech recognition software

(Hinton *et al.*, 2012) and they have mastered the ancient board game of Go (Silver *et al.*, 2016). In the context of geolocation, Weyand *et al.* (2016) have developed the so-called PlaNet method that is capable of geolocating photographs taken across the globe with accuracy that is superior to the average human. PlaNet partitions the spatial domain into discrete cells and fits a convolutional neural network to millions of images taken from these cells. A similar approach could prove fruitful for capturing the complexity that is inherent in high dimensional microbiome data, but the approach must be modified to produce point level geolocation predictions.

We propose DeepSpace: a novel forensic geolocation algorithm which merges features of spatial point pattern theory and deep learning to model spatially distributed point level data. In particular, DeepSpace estimates the intensity surface of a spatial point pattern by using an ensemble of deep neural network classifiers trained on random Voronoi partitions of the spatial domain. By fitting a flexible deep neural network classifier to each partition and averaging over partitions, the DeepSpace model can approximate the continuous conditional distribution of a sample's provenance given its microbiome composition. Indeed, as the number of cells within a partition increases, the discrete approximation to a continuous probability density function can be made arbitrarily precise, and the universal approximation property of neural networks (Goodfellow *et al.*, 2016) ensures that the cell probabilities fall in the span of the neural network classifier.

We use DeepSpace to analyse the dust microbiome data at the national (continental USA), regional (North Carolina's tricounty research triangle) and global (nearly 30 countries across six continents) levels. An objective of this analysis is to determine the scales at which geolocation using microbiome data is feasible. We compare DeepSpace with other geolocation algorithms and find a substantial reduction in cross-validation error at the national level, and for the first time we demonstrate that microbiome data can be used to geolocate dust samples at the global level successfully.

2. Data

2.1. National and regional data

A complete description of the data collection and processing procedures can be found in Barberán *et al.* (2015a, b) and Grantham *et al.* (2015) and so we describe these procedures only briefly. The data were collected by citizen science volunteers as part of the 'Wild life of our homes' (WLOH) project (`homes.yourwildlife.org`). We analysed dust samples that were taken by the volunteers from the upper door trim of the outside surface of an exterior door. This data set includes $n = 1301$ dust samples harbouring $p = 57331$ distinct fungal phylotypes identified with high throughput sequencing. We also performed a separate analysis of the subset of the samples from Wake, Durham and Orange counties in central North Carolina ($n = 116$).

2.2. Global data

In addition to the WLOH analysis, we conduct the first analysis of dust-associated microorganisms that have been collected from nearly 30 countries across six continents (excluding Antarctica): the largest database of global dust samples to date. 10–15 samples were collected from each country representing several geographic regions: the Americas (Mexico, Costa Rica, Colombia, Trinidad and Tobago, Uruguay and Argentina), Africa (Ghana, Nigeria and South Africa), eastern Europe (Czechia, Croatia, Hungary and Macedonia (the former Yugoslav Republic of Macedonia)), western Asia (Turkey (sample collection in Turkey was split evenly between the cities of Istanbul and Antalya), Cyprus (specifically, samples were collected from the northern part of the island of Cyprus, which, although recognized as belonging to the Republic of Cyprus, is *de facto* controlled by the self-declared Turkish Republic of Northern Cyprus)

and Jordan), the Middle East (Kuwait, Qatar, Oman, Georgia and Azerbaijan), central Asia (Kazakhstan and Pakistan), eastern Asia (Vietnam, South Korea and Malaysia) and Oceania (Australia and New Zealand). Details of the sample collection, preparation and sequencing are described in Appendix A. In total, the pipeline yielded $p = 15475$ distinct fungal phylotypes across $n = 399$ samples.

3. Statistical methods

Geolocation may be formalized in the language of spatial point process theory (Baddeley *et al.*, 2015; Gelfand *et al.*, 2010) where the response is a spatial location $\mathbf{s} \in \mathcal{D}$, which is regressed onto non-spatial covariate information $\mathbf{x} \in \mathcal{X}$. We assume that the spatial point process follows a non-homogeneous Poisson process with intensity surface $\lambda(\mathbf{s}|\mathbf{x})$ for all $\mathbf{s} \in \mathcal{D}$. The challenge lies in choosing an appropriate model for the intensity surface $\lambda(\mathbf{s}|\mathbf{x})$.

One potential approach is to model $\lambda(\mathbf{s}|\mathbf{x})$ as a log-Gaussian process, placing the point process in the well-known Cox process family (Møller and Waagepetersen, 2003). This choice complicates the evaluation of the point process likelihood, however, as it includes an integral over $\mathcal{D}$ which is stochastic with regard to the Gaussian process and must be approximated numerically. Liang *et al.* (2008) formulated a Monte Carlo algorithm using a knot-based predictive process approximation predicated on careful knot selection within the spatial domain.

In our application, $\mathcal{D}$ is continental USA and every geocoded location $\mathbf{s}$ yields microbial presence or absence data $\mathbf{x}$ from $\mathcal{X} = \{0, 1\}^p$ with p in the tens of thousands. Implementing the aforementioned Monte Carlo algorithm may prove problematic when the spatial locations exhibit strong clustering over the extent of $\mathcal{D}$ and/or the non-spatial covariate space is high dimensional. To accommodate these features of our data better, we develop a new algorithm for estimating the intensity surface of a spatial point process which relies on ensembles of supervised classifiers to learn spatial variation patterns.

3.1. DeepSpace algorithm

Let $\mathcal{P} = \{P_k\}_{k=1}^K$ denote a partition of $\mathcal{D}$. Given $\mathcal{P}$, we assume that the Poisson intensity is constant in each cell of this partition, and we denote $\lambda_k(\mathbf{x})$ as the intensity for all $\mathbf{s} \in P_k$. Operating at this discrete level offers numerous benefits. First, it allows us to avoid formulating the intensity as a log-Gaussian process realization, in turn avoiding the difficulties that are associated with likelihood evaluation. Second, with the intensity surface discretized, its estimation is reframed as a supervised classification problem where regularization or dimension reduction techniques have been developed and studied extensively (Friedman *et al.*, 2001). Finally, fitting a piecewise constant intensity forces parsimony and thus has a regularizing effect over space. As described below, by averaging over uncertainty in $\mathcal{P}$ and aggregating these piecewise constant intensities we can approximate a continuous intensity function and capture complex dependence between the covariates and the intensity surface.

Let $|P_k|$ denote the area of region P_k and take $f_k(\mathbf{x}) = \log\{|P_k|\lambda_k(\mathbf{x})\}$ and the probability on region k is

$$\Pr(\mathbf{s} \in P_k|\mathbf{x}) = \frac{\exp\{f_k(\mathbf{x})\}}{\sum_{l=1}^K \exp\{f_l(\mathbf{x})\}} \qquad \text{for } k = 1, \ldots, K. \tag{1}$$

We estimate $f_1, \ldots, f_K$ by training a supervised classifier on available data points as follows. Define $h(\mathbf{s}) = k$ for $s \in P_k$, which serves as the label for the cell to which $\mathbf{s}$ belongs; then the classifier is trained on $\{(\mathbf{x}_i, h_j(\mathbf{s}_i))\}_{i=1}^n$. The trained classifier yields estimators $\hat{f}_1, \ldots, \hat{f}_K$ and subsequently the intensity estimator $\hat{\lambda}_k(\mathbf{x}) = |P_k|^{-1} \exp\{\hat{f}_k(\mathbf{x})\}$.

The spatial prediction is the location that maximizes the fitted intensity surface. A piecewise constant intensity surface does not have a unique maximum, and thus the precision of our geolocator is limited by the resolution of partition $\mathcal{P}$. Thus, to obtain finer accuracy we propose to construct multiple random partitions of $\mathcal{D}$, to train a separate classifier on each partition and to estimate the intensity surface as the average over the random partitions. To generate partition $j = 1, \ldots, J$, we draw K_j 'seeds' from the domain, $\mathbf{v}_{jk} \sim^{\text{IID}} \text{uniform}(\mathcal{D})$ and define a Voronoi partition $\mathcal{P}_j = \{P_{jk}\}_{k=1}^{K_j}$ where $P_{jk} = \{\mathbf{s} \in \mathcal{D} : \|\mathbf{s} - \mathbf{v}_{jk}\| < \|\mathbf{s} - \mathbf{v}_{jl}\| \text{ for } l \neq k\}$. We train a supervised classifier on $\{[\mathbf{x}_i, h_j(\mathbf{s}_i)]\}_{i=1}^n$ to yield $\hat{F}_j = \{\hat{f}_{j1}, \ldots, \hat{f}_{jK_j}\}$ and subsequently $\hat{\lambda}_{jk}(\mathbf{x}) = |P_{jk}|^{-1} \exp\{\hat{f}_{jk}(\mathbf{x})\}$.

We obtain geolocation predictions by averaging pointwise over our collection of models $\mathcal{M} = \{\hat{F}_j\}_{j=1}^J$. The estimated Poisson intensity is $\hat{\lambda}(\mathbf{s}|\mathbf{x}) = J^{-1} \Sigma_{j=1}^J \hat{\lambda}_{jh_j(\mathbf{s})}(\mathbf{x})$. For a sample of unknown origin with non-spatial covariate information $\mathbf{x}$, a location $\mathbf{s}$ is attributed a geolocation score given by

$$g(\mathbf{s}|\mathbf{x}, \mathcal{M}) = \frac{\hat{\lambda}(\mathbf{s}|\mathbf{x})}{\int \hat{\lambda}(\mathbf{s}|\mathbf{x}) d\mathbf{s}}. \tag{2}$$

Thus, the estimated location for a sample with microbiome composition $\mathbf{x}$ is

$$\hat{\mathbf{s}} = \underset{\mathbf{s} \in \mathcal{D}}{\arg\max}\ g(\mathbf{s}|\mathbf{x}, \mathcal{M}) \tag{3}$$

and $1 - \alpha$ prediction regions are computed as $R_\alpha = \{\mathbf{s} : g(\mathbf{s}|\mathbf{x}, \mathcal{M}) > T_\alpha\}$ for the threshold T_α, which gives $\int I\{g(\mathbf{s}|\mathbf{x}, \mathcal{M}) > T_\alpha\} g(\mathbf{s}|\mathbf{x}, \mathcal{M}) d\mathbf{s} \approx \alpha$. In practice, computing the estimated location and prediction interval is performed by evaluating the predictive distribution at a fine grid of points covering the spatial domain.

The accuracy of these predictions depends heavily on the performance of the trained classifiers. The user is free to select any supervised classifier to fit to their spatially distributed data, so long as the classifier is designed to receive the input data and returns an array of predicted class probabilities summing to 1. At present, popular choices include classical machine learning classifiers (Friedman *et al.*, 2001) such as K nearest neighbours, support vector machines and random forests, deep learning classifiers (Goodfellow *et al.*, 2016), such as deep neural networks and convolutional neural networks (for image data), recurrent neural networks (for sequential data), or an ensemble of classifiers, where a classifier is selected at random to fit to each new Voronoi partition. Thus, the algorithm accommodates a vast number of possible spatial prediction applications, and the user should choose from classifiers that are appropriate for their data-generating model.

Deep learning is well suited for the high dimensionality and complex dependence structure of our motivating microbial presence or absence data. When paired with a deep learning classifier, such as the deep neural network that is described below, we call our proposed algorithm DeepSpace for its ability to learn deep patterns in spatially distributed data. DeepSpace spans a rich class of intensity functions $\lambda(\mathbf{s}|\mathbf{x})$ if we select a suitable classifier for the relationship between the covariates and the aggregate probabilities of the K regions that are defined by a single partition. Although each estimated intensity function $\hat{\lambda}_j(\mathbf{s}|\mathbf{x})$ is discontinuous in space, the model-averaged geolocation function in equation (2) can approximate a continuous intensity function as J increases. Furthermore, if we select a classifier which consistently estimates the aggregate probabilities for the K regions that are defined by a single partition, then the model-averaged estimator (2) should be a consistent non-parametric estimator for the true geolocation function $g(\mathbf{s}|\mathbf{x})$.

3.2. Deep neural network classifier

A neural network (Goodfellow *et al.*, 2016) is comprised of p input nodes (input layer), K output nodes (output layer), one for each available class, and R 'hidden' nodes that connect the nodes of the input layer with those of the output layer. The nodes in the hidden layer are called neurons, as their non-linear transformations of the data are designed to mimic the firing, or 'activation', of neurons in the brain when processing sensory data. The single-layer neural network for $k = 1, \ldots, K$ is

$$f_k(\mathbf{x}) = \beta_{k0} + \sum_{r=1}^{R} \beta_{kr}\sigma(z_r) \qquad z_r = \alpha_{r0} + \sum_{j=1}^{p} \alpha_{rj}x_j, \tag{4}$$

$\sigma(\cdot)$ is a non-linear function, typically chosen to be the inverse logistic function $\sigma(z) = \{1 + \exp(-z)\}^{-1}$ or the rectified linear unit function $\sigma(z) = \max\{0, z\}$, $\boldsymbol{\alpha}_r = (\alpha_{r0}, \ldots, \alpha_{rp})'$ are unknown coefficients from the regression of neuron z_r on $\mathbf{x}$ with intercept $r = 1, \ldots, R$, and $\boldsymbol{\beta}_k = (\beta_{k0}, \ldots, \alpha_{kR})'$ are unknown coefficients from regressing $f_k(\mathbf{x})$ on activations $\sigma(z_1), \ldots, \sigma(z_R)$ with intercept, $k = 1, \ldots, K$.

Until recently, attempts to introduce more than one hidden layer into a neural network were fraught with convergence difficulties. With the advent of 'pretraining' methods designed to alleviate these barriers (Hinton *et al.*, 2006), however, neural networks experienced a renaissance in the form of 'deep learning', so named for their new found ability to learn much richer associations in the data than is possible with a single hidden layer (LeCun *et al.*, 2015). The single-layer neural network (4) is generalized to a deep neural network by introducing hidden layers $l = 1, \ldots, L$ each with R_l neurons so that

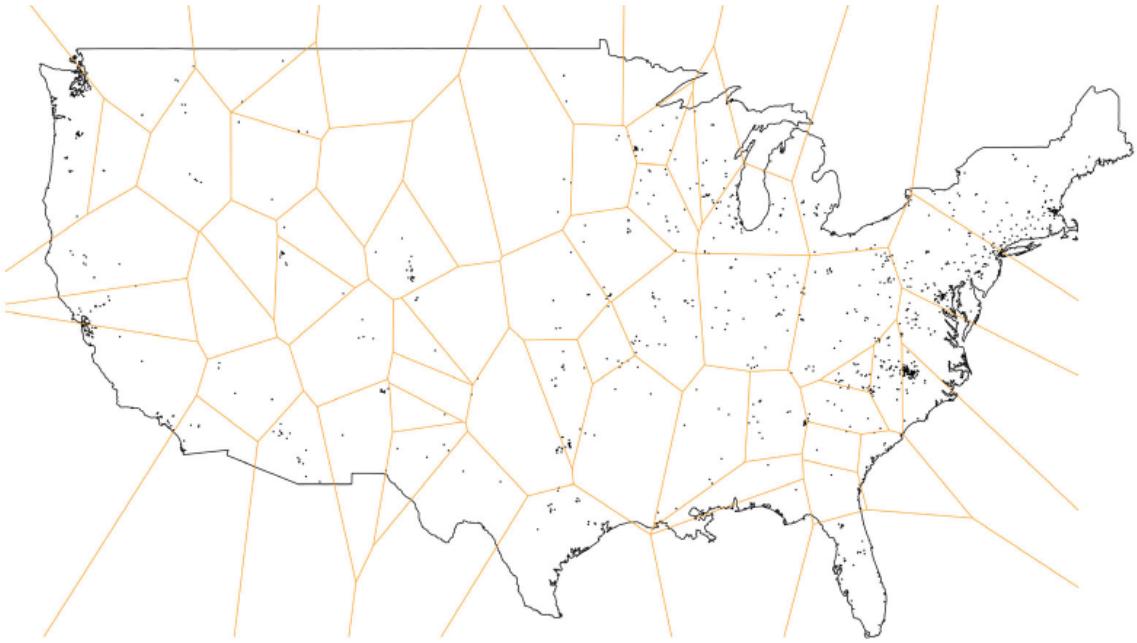
$$f_k(\mathbf{x}) = \beta_{k0} + \sum_{r=1}^{R_L} \beta_{kr}\sigma(z_r^L) \qquad z_r^l = \begin{cases} \alpha_{r0}^l + \sum_{s=1}^{R_{l-1}} \alpha_{rs}^l \sigma(z_s^{l-1}) & l = 2, \ldots, L, \\ \alpha_{r0}^1 + \sum_{j=1}^{p} \alpha_{rj}^1 x_j & l = 1. \end{cases}$$

To arrive at equation (1), we apply the multiple logistic function, which is otherwise known as the softmax, to $f_1, \ldots, f_K$ of the final output layer. Furthermore, as each classifier–partition pair of DeepSpace aims to determine the probability of each region $k = 1, \ldots, K$ producing a sample with covariate data $\mathbf{x}$, we select the categorical cross-entropy function to measure the error of our network, which penalizes the network for assigning high probabilities to incorrect regions. Finally, the back-propagation algorithm and a stochastic gradient-based optimization method, such as Adam (Kingma and Ba, 2014) are used to obtain estimated $\hat{f}_1, \ldots, \hat{f}_K$ which minimize the cross-entropy cost over several training epochs.

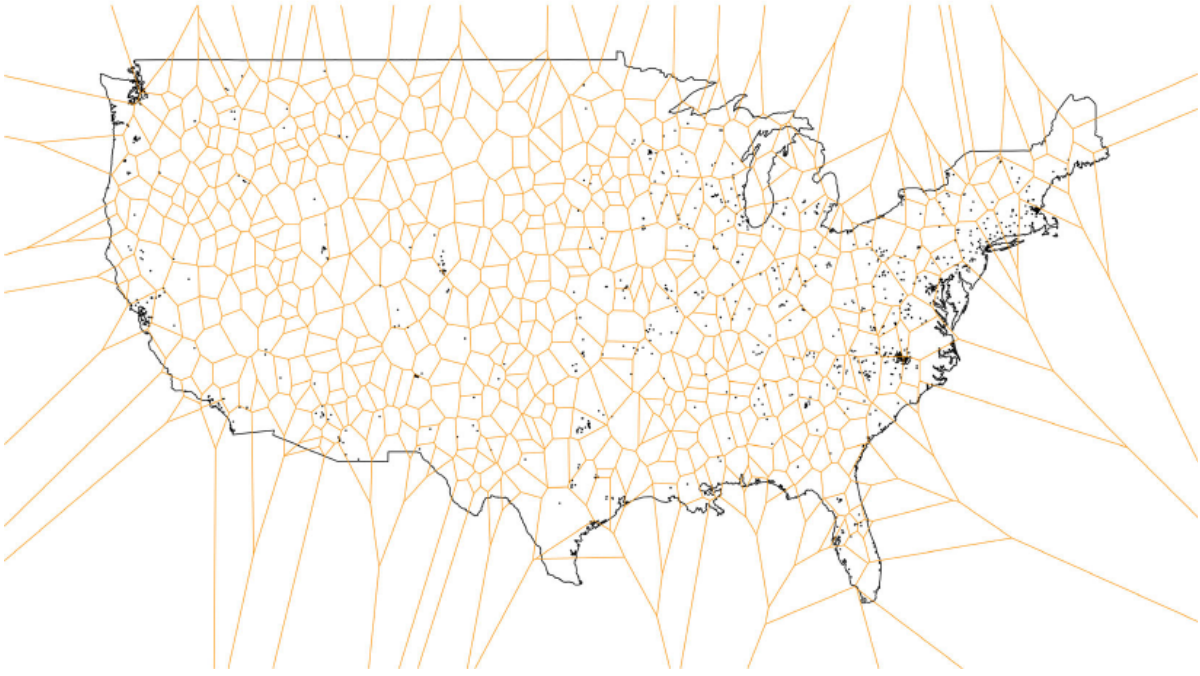
3.3. Tuning

In addition to the tuning parameters in a standard neural network, such as the number of hidden layers and the number of neurons per layer, DeepSpace requires selecting the number of partition seeds and the number of random partitions. As will be demonstrated in Section 4, cross-validation is a reliable albeit expensive means for arriving at appropriate settings for these tuning parameters.

The number of seeds, K, affects the granularity of the Voronoi partitions and ought to be chosen on the basis of the number of sample points, n, and their spatial distribution over the domain, $\mathcal{D}$. If K is small, then the partitioning is coarse (Fig. 1(a)) and most cells are suitably populated, but they may not discriminate well between different regions in $\mathcal{D}$ and the classifier will infer little about the true spatial variation of the data. If K is large, the partitioning is fine



(a)



(b)

Fig. 1. Voronoi partitions (——) generated over continental USA (sample points (●) are labelled according to the cell that they lie within and a supervised classifier is trained on these labelled data): (a) coarse partition with the number of cells chosen to be approximately 5% of the number of sample points; (b) fine partition with the number of cells chosen to be approximately 50% of the number of sample points

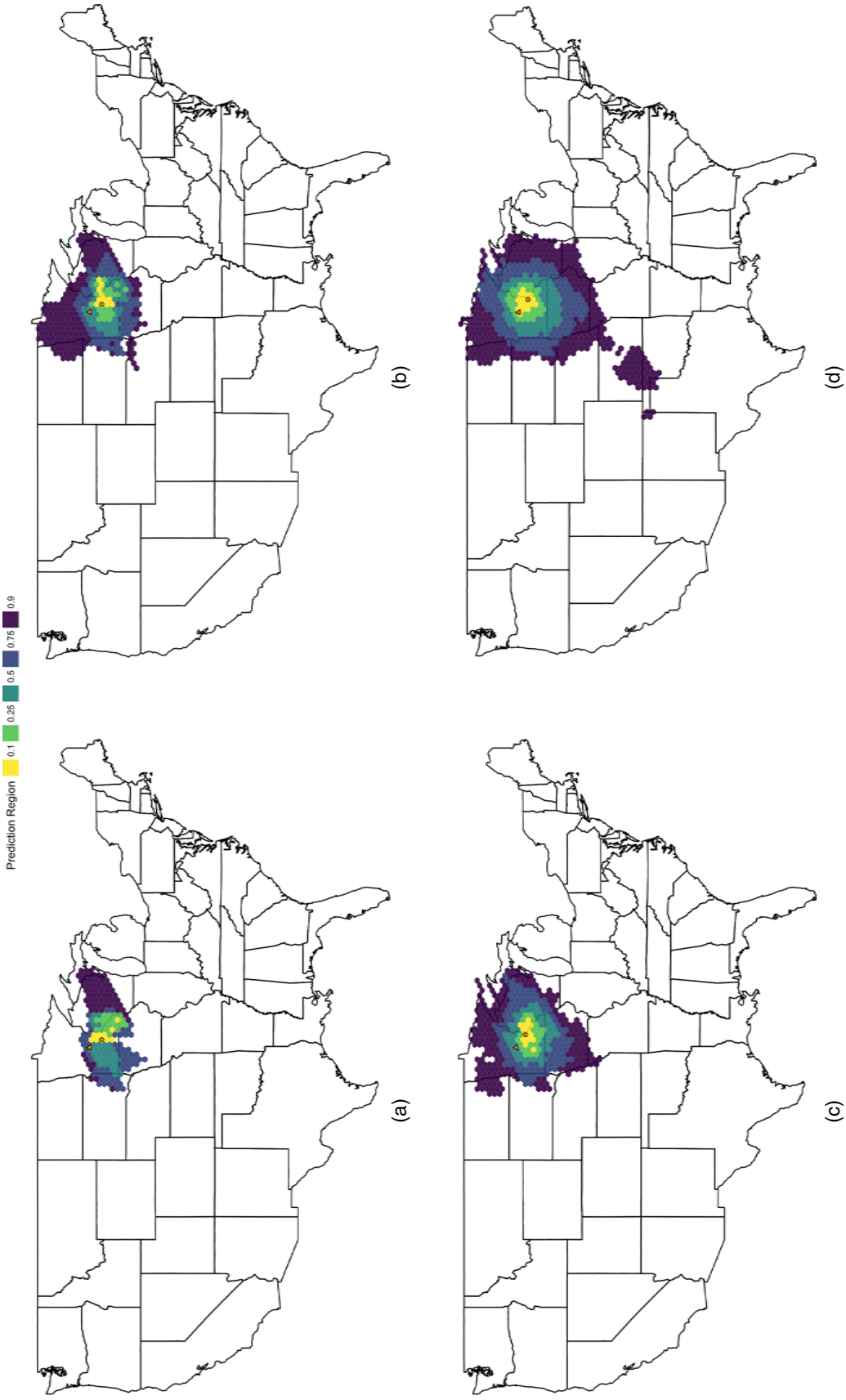


Fig. 2. Pointwise and regional predictions for a single sample made by coarse DeepSpace at increasing numbers of classifier–partition pairs (○, predicted; △, true): (a) five pairs; (b) 10 pairs; (c) 20 pairs; (d) 50 pairs

(Fig. 1(b)) and most cells lack a suitable number of representative samples for the classifier to learn useful representations of the data from these regions, though it may detect fine scale variation in well-populated regions. Of course, there is no need to fix the number of seeds for every partition, and one may vary the seed number at the creation of each partition. We explore the effects of these three partitioning schemes (coarse, fine and mixed) on geolocation performance in Section 4.

DeepSpace depends on a large number of partition–classifier pairs J to perform effectively. Although there is no harm in using too many pairs, there is a point of diminishing returns where successive pairs do not measurably improve predictions and are therefore computationally burdensome. It is possible to arrive at a suitable number of partitions by withholding a subset of one's data and measuring the change in prediction performance with each successive partition–classifier pair added. In particular, one can track changes in prediction error (the distance between each sample's true and predicted origin) and choose to suspend fitting further partition–classifiers when errors have stabilized. Fig. 2 portrays pointwise and regional predictions that were made by DeepSpace with $J = 5$, 10, 20, 50. For this single sample, the prediction region widens with J, including a low probability region stretching into Oklahoma. However, the predictions have relatively stabilized on the upper Midwest near its true origin, suggesting that including more pairs may add little value. Of course, this determination ought to be made after viewing the model's prediction performance for a subset of available samples.

3.4. Computation

DeepSpace is written in Python 3. Deep neural networks are implemented by using Keras (Chollet, 2015): a popular deep learning library that is designed to interface with numerical computing back-ends, including TensorFlow and Theano. Code is available from `https://github.com/nsgrantham/forensic-geolocation`.

4. Analysis

We demonstrate DeepSpace on the dust-associated fungal microbiome data sets that were described in Section 2. In particular, we analyse the WLOH data at the national level, with samples distributed over continental USA, as well as a subset of these data at the regional level, with samples across the research triangle of central North Carolina, encompassing the cities of Raleigh, Durham and Chapel Hill. At the global level, we analyse a new data set of dust-associated fungal microbiome samples from countries across eastern Europe, the Middle East, Africa, Asia, Oceania and the Americas.

We compare the following models:

(a) *Spatial NN*, a geolocation algorithm using K-nearest-neighbours classifiers with 25 neighbours and the Jaccard distance metric that is suitable for binary data;
(b) *Spatial RF*, a geolocation algorithm using random-forests classifiers with 200 estimators;
(c) *Spatial Net*, a geolocation algorithm using neural network classifiers with one hidden layer of 2048 neurons each with a rectified linear unit activation function (the output layer is given a softmax activation function with cross-entropy cost and optimization is performed with Adam (Kingma and Ba, 2014) using an initial learning rate of 0.001 and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$);
(d) *DeepSpace*, a geolocation algorithm using deep neural network classifiers with three hidden layers of 2048, 1024 and 1024 neurons, each with the rectified linear unit activation function (a dropout rate of 0.3 is placed between the layers of the network to prevent

overfitting; the output layer is given a softmax activation function with cross-entropy cost and optimization is performed with Adam (Kingma and Ba, 2014) using an initial learning rate of 0.001 and parameters $\beta_1 = 0.9$ and $\beta_2 = 0.999$);

(e) *BDA*, the geolocation model that was developed in Grantham *et al.* (2015);
(f) *Area DNN*, a deep neural network classifier built as in DeepSpace, but not placed within the geolocation algorithm; rather, this deep neural network predicts a sample point's area of origin, i.e. state (national), county (local) and country (global), and uses the predicted area's population centroid as the 'most likely' origin.

We further compare several partitioning schemes, including

(a) *coarse*—K is fixed at approximately 5% of n (e.g. Fig. 1(a)),
(b) *fine*—K is fixed at approximately 50% of n (e.g. Fig. 1(b)),
(c) *mixed*—K is allowed to vary and is chosen at random to be 5%, 10%, ..., 45%, 50% of n for each new partition, and
(d) *none*—no seeds are selected (suitable for the BDA and Area DNN models).

We perform cross-validation to test the performance of these models under the available partitioning schemes at the national, regional and global levels. We randomly partition samples into 10 folds and train the models 10 times, each time excluding one of the 10 folds in the training set. After fitting a model to the training data, we use the model to geolocate samples from the withheld fold and calculate prediction errors (the great circle distance between a sample's true and predicted origin), prediction region coverage (the proportion of samples that are captured within a model's $1-\alpha$ probability region) and prediction area matching (the proportion of predictions that lie in the same city or locality, county, if applicable, state or foreign equivalent, and country as that of the true sample point).

4.1. National

We begin with an analysis of the WLOH data set with $n = 1301$ samples of $p = 57331$ distinct fungal phylotypes. Under the assumption that the sampling density should reflect the population density across the USA, we weight samples upwards or downwards from states where the number of samples respectively underrepresents or overrepresents the state population relative to the national population. For instance, North Carolina contributes the most samples despite being the ninth most populous state, so the contribution of each individual North Carolina point to the model is downweighted to reflect density patterns nationwide better. Specifically, let A_t, $t = 1, \ldots, T$, represent the population areas (in this case continental USA states, $T = 48$) and $\text{pop}(A_t)$ be the population of A_t, and define $A(\mathbf{s}) = A_t$ for $\mathbf{s} \in A_t$. Then a sample point with origin $\mathbf{s}_i$ is given weight $w_i = [\text{pop}\{A(\mathbf{s}_i)\}/\Sigma_{t=1}^{T} \text{pop}(A_t)][\Sigma_{j=1}^{n} I\{\mathbf{s}_j \in A(\mathbf{s}_i)\}]^{-1}$.

Table 1 summarizes the performance of the competing geolocation models. For all three partitioning schemes, there is a reduction in prediction error as the complexity of the spatial classifier increases (Spatial NN to Spatial RF to Spatial Net to DeepSpace). The coarse partitioning scheme appears to lead consistently to the lowest prediction error but, because of the large size of the cells, the coverage probabilities are overinflated. The fine partitioning achieves slightly lower coverage probabilities at the expense of higher prediction errors. As might be expected, mixed partitioning strikes a balance between coarse and fine with respect to their differences in prediction error and probability region coverage. Among the spatial models, DeepSpace regularly outperforms Spatial NN, Spatial RF and Spatial Net in predicting a sample's true state, county and city of origin. In fact, DeepSpace appears better at predicting the state of origin for a sample than a deep neural network fit specifically for state classification (Fig. 3). Thus,

Table 1. Geolocation predictions within continental USA by using ambient dust-associated fungal microbiome data†

Seeds	*Model*	*ME (km)*	*COV*	*Area match (%)*		
				State	*County*	*City*
Coarse	Spatial NN	231.9	96.0	44.9	12.6	10.4
	Spatial RF	194.8	99.2	48.3	13.9	11.9
	Spatial Net	87.7	97.0	60.6	22.1	17.6
	DeepSpace	86.9	98.2	61.0	21.5	17.2
Fine	Spatial NN	258.3	80.5	43.9	14.5	11.8
	Spatial RF	221.5	96.5	46.8	17.2	14.8
	Spatial Net	119.5	89.0	56.5	24.4	20.7
	DeepSpace	107.6	93.1	58.7	23.8	20.0
Mixed	Spatial NN	247.9	90.0	44.6	14.6	12.1
	Spatial RF	213.7	98.6	47.6	17.0	14.2
	Spatial Net	113.3	94.3	58.2	23.9	19.7
	DeepSpace	97.8	96.3	60.2	23.6	19.4
None	BDA	263.7	91.0	31.9	1.6	0.8
	State Area DNN	203.9	—	57.0	—	—

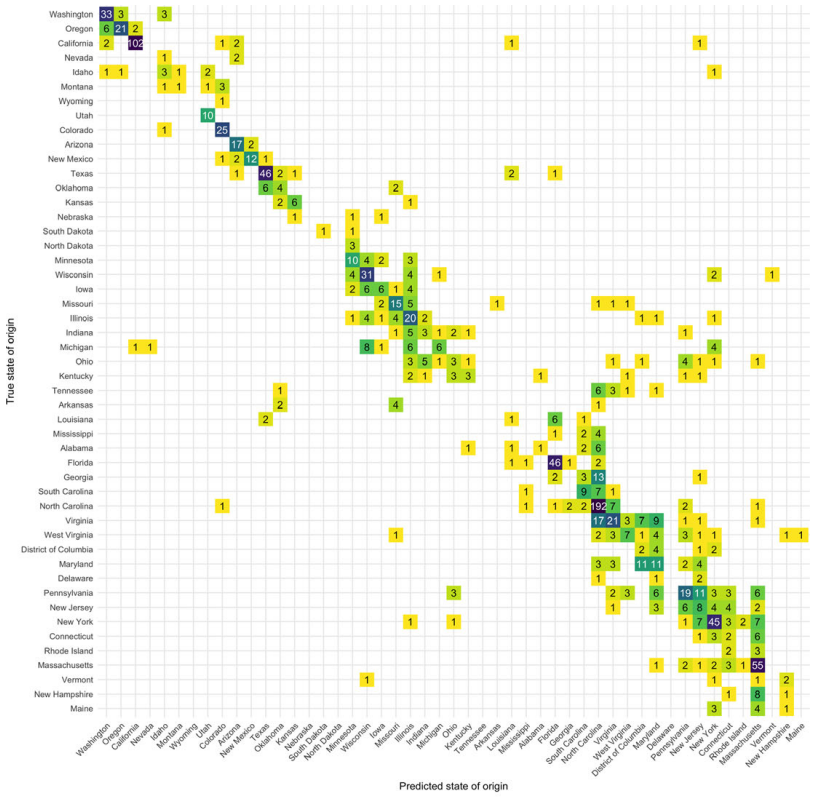
†Methods are compared in terms of median absolute prediction error ME, coverage COV of 90% prediction regions and classification accuracy for predicting the state, county and city.

operating at the point level appears beneficial, perhaps because it enables DeepSpace to learn regional patterns that help it to distinguish between states.

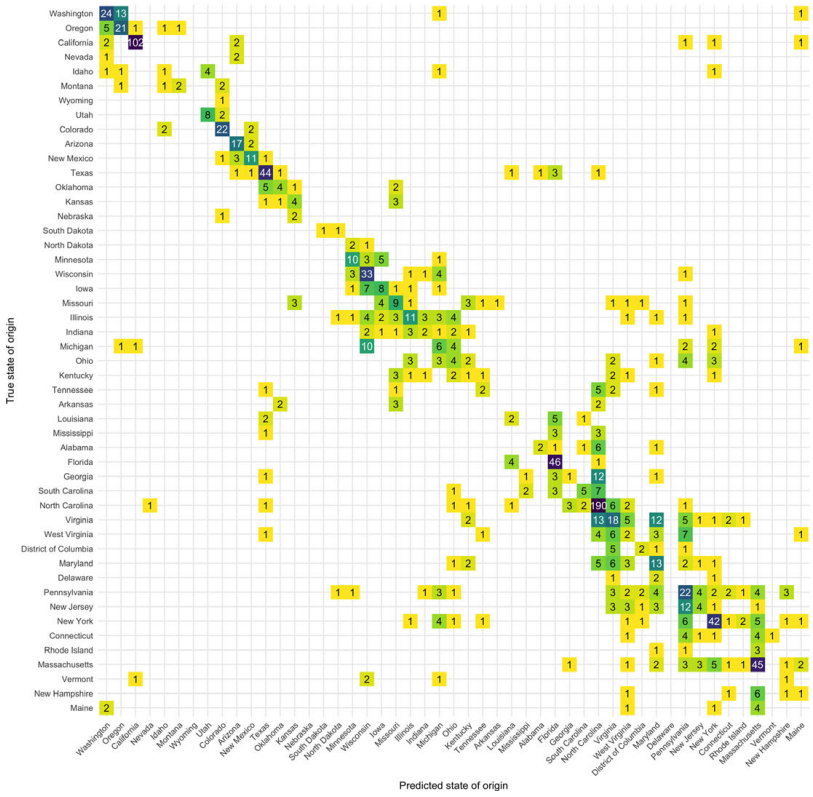
Figs 4 and 5 depict the average prediction errors that are made by Spatial RF and DeepSpace respectively, under coarse partitioning. In these plots, the arrows point to the average prediction direction and the colour indicates the average prediction error. The averages are computed by using a Gaussian kernel smoother with a bandwidth of 100 km. Both models are biased towards populated urban areas for which more data are available than for surrounding rural areas. Spatial RF does not appear to detect quite as many regional patterns as does DeepSpace, where the latter model identifies many more areas with low prediction error throughout the Northeast, Midwest and Western USA. Despite downweighting points in North Carolina to reflect population patterns better nation wide, central North Carolina attracts a large number of predictions away from samples that were collected in Virginia, Tennessee and the southern USA (with the exception of Florida). The results in North Carolina are fairly consistent with the general trend of biased predictions towards urban areas: a reflection of the sampling scheme that is meant to be representative of population density.

4.2. Regional

We now analyse a subset of the WLOH data: only $n = 116$ samples with $p = 20557$ distinct fungal phylotypes belonging to the central North Carolina counties of Wake, Durham and Orange. At the national scale, differences in fungal occupancy are likely to reflect both biogeographical differences in terms of which taxa occur where and local habitat differences. When we perform analysis at this regional scale, the taxa pool of fungi that could colonize any particular site are essentially the same across all sites and differences are more likely to be due to which taxa can survive in local habitats. By focusing on a small geographic area we can isolate the ability of the models to predict the origin of a sample when biogeographic differences are held relatively



(a)



(b)

Fig. 3. State classification between (a) coarse DeepSpace and (b) state Area DNN: each plot depicts the frequency of samples predicted to originate from a state (*x*-axis) against their true state of origin (*y*-axis)

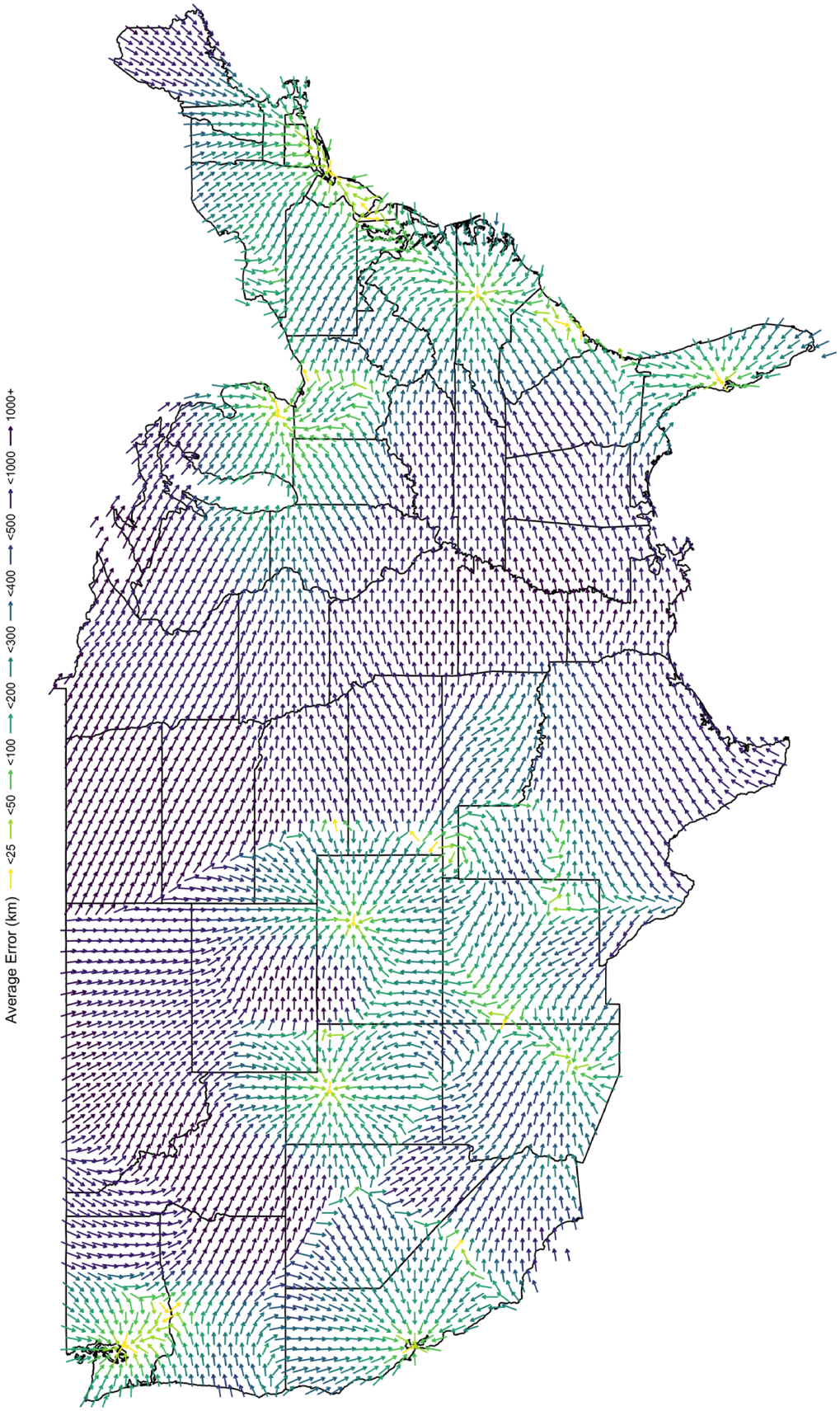


Fig. 4. Average prediction errors made by coarse Spatial RF over tenfold cross-validation: each arrow points from a sample location in the direction to which a prediction is likely to be made, and the arrow's colour indicates how far on average that prediction is likely to be from the true origin

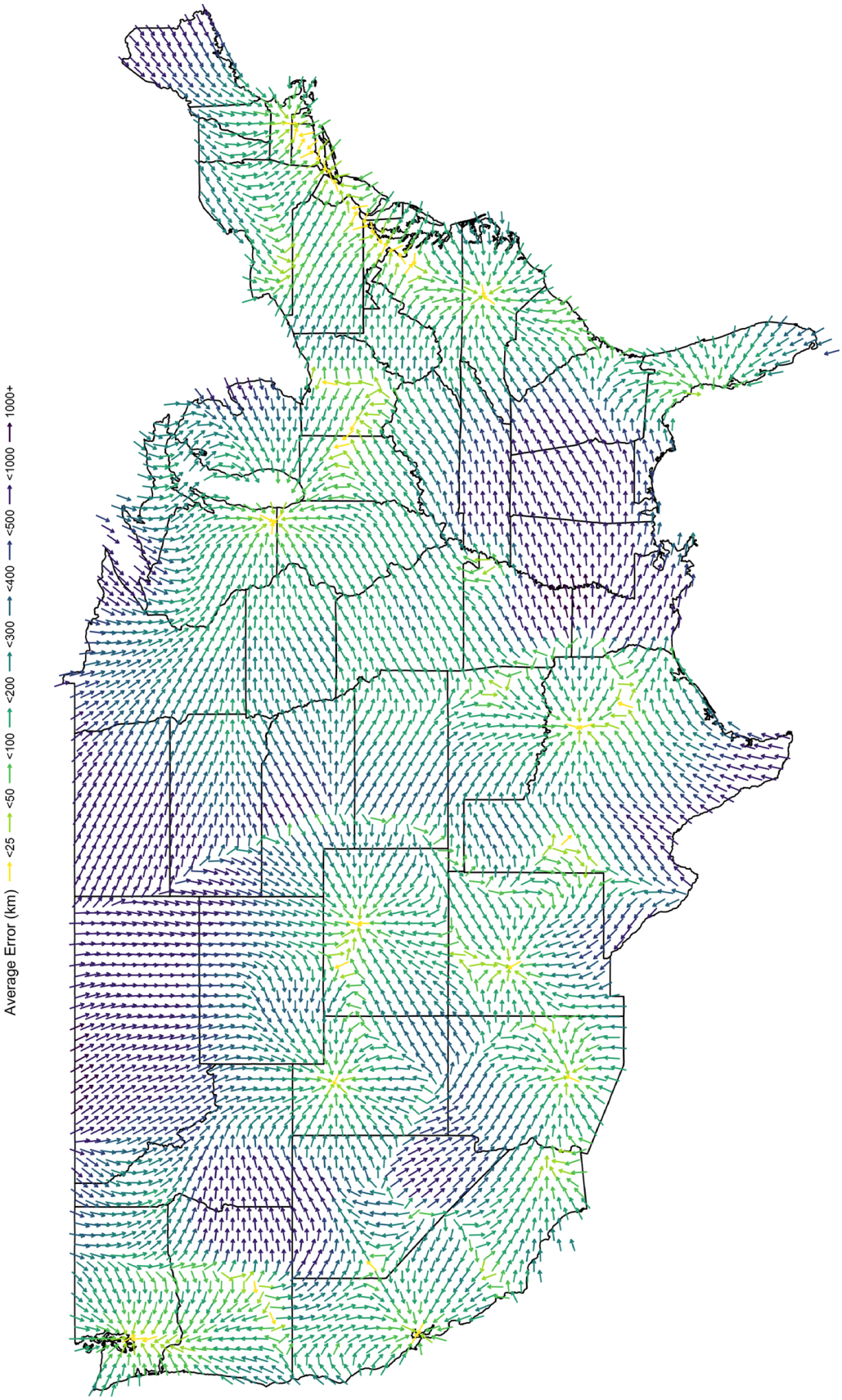


Fig. 5. Average prediction errors made by coarse DeepSpace over tenfold cross-validation: each arrow points from a sample location in the direction to which a prediction is likely to be made, and the arrow's colour indicates how far on average that prediction is likely to be from the true origin

Table 2. Geolocation predictions within the triangle region of North Carolina by using ambient dust-associated fungal microbiome data†

Seeds	*Model*	*ME (km)*	*COV*	*Area match (%)*	
				County	*City*
Coarse	Spatial NN	19.2	91.4	44.8	25.0
	Spatial RF	17.4	98.3	38.8	29.3
	Spatial Net	18.9	90.5	41.4	21.6
	DeepSpace	19.1	89.7	44.0	24.1
Fine	Spatial NN	23.3	73.3	40.5	17.2
	Spatial RF	19.5	87.9	36.2	16.4
	Spatial Net	18.5	76.7	40.5	17.2
	DeepSpace	19.6	82.8	40.5	18.1
Mixed	Spatial NN	20.4	84.5	43.1	24.1
	Spatial RF	20.2	93.1	36.2	19.0
	Spatial Net	19.2	90.5	49.1	25.9
	DeepSpace	20.0	90.5	40.5	18.1
None	BDA	19.5	90.5	40.5	19.0
	County Area DNN	18.0	—	53.4	—

†Methods are compared in terms of median absolute prediction error ME, coverage COV of 90% prediction regions and classification accuracy for predicting the county and city.

constant. In this analysis we seek to determine whether there is a limit to the resolution that we may geolocate samples by using fungal occupancy data.

Table 2 summarizes geolocation efforts over this localized region. As in the national level analysis, we weight samples proportionally to population density, this time by county population. Unlike the national level analysis, many of the spatial models perform similarly across the three partitioning schemes. In fact, BDA and county Area DNN achieve the lowest prediction errors. Moreover, county Area DNN has the highest county classification accuracy (53.4%) compared with the spatial models (from 36.2% to 49.1%). The county classification rates for all models exceeds $\frac{1}{3}$, suggesting that there is some hope of county level prediction with these data. Fig. 6 shows average prediction errors made by DeepSpace and county Area DNN within the triangle (using a smoother bandwidth of 5 km), where the former makes most of its predictions between Durham and Raleigh, and the latter centres on Raleigh, which is the most populous city in the area.

4.3. Global

Finally, we analyse a global level dust-associated microbiome data set with $n = 399$ samples with $p = 15475$ distinct fungal phylotypes collected across countries in eastern Europe, the Middle East, Africa, Asia, Oceania and the Americas. There were 10–15 sampling locations in each country. Because of this relative balance in sample collection, we do not weight the samples to reflect country population. Moreover, samples within each country often stem from a single major city and thus do not permit estimation of within-country variation. We therefore compare the models only on their ability to detect a sample's country of origin. As was demonstrated in the national level analysis, DeepSpace is capable of learning regional patterns in the data, which could improve its classification accuracy.

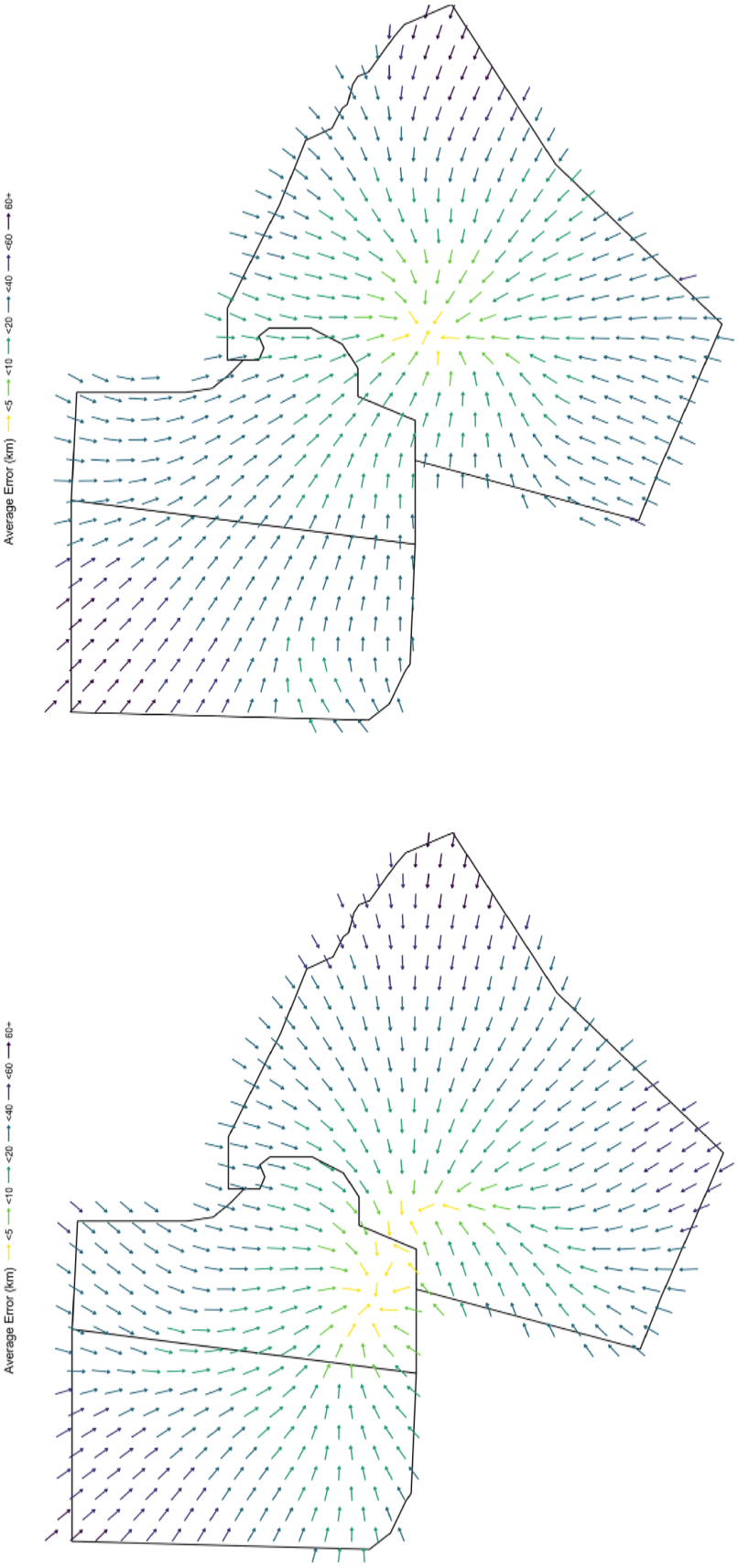
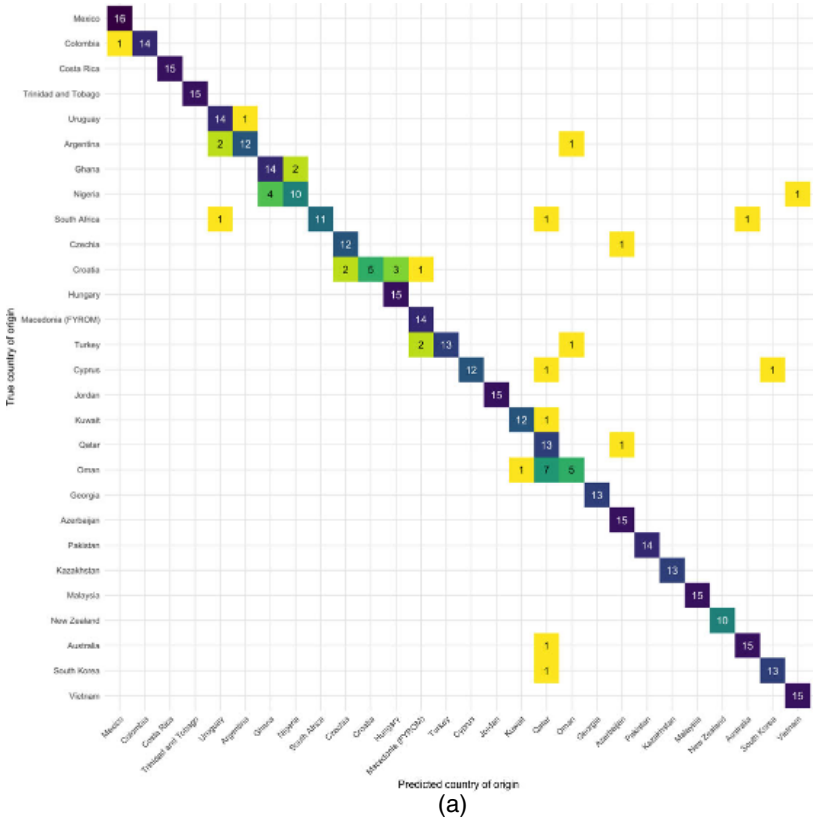
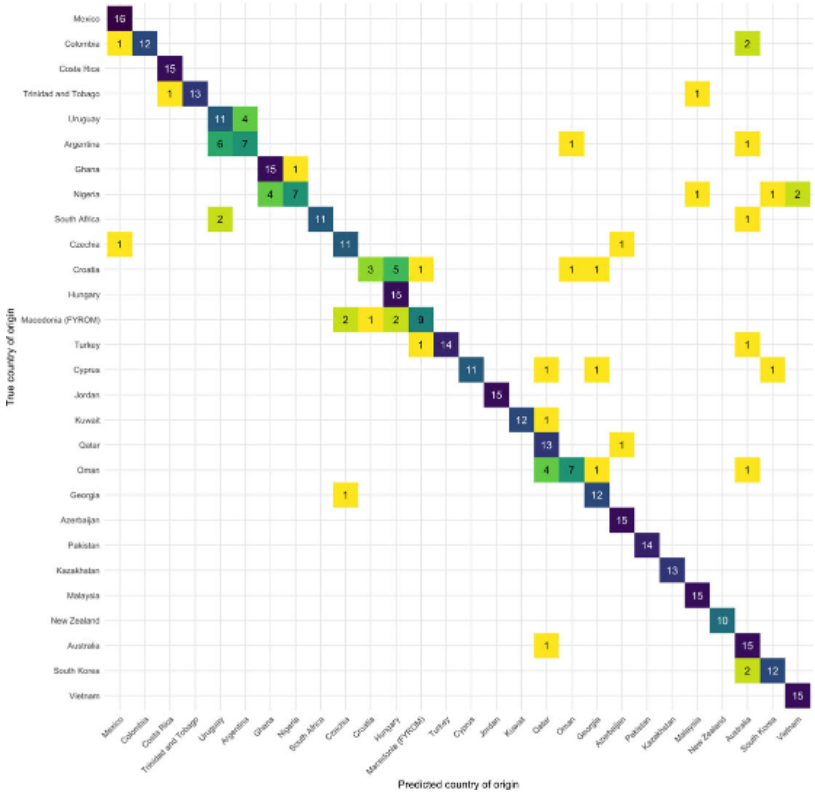


Fig. 6. Prediction errors made by (a) coarse DeepSpace and (b) county Area DNN over tenfold cross-validation: each arrow points in the direction to which a prediction is likely to be made, and the arrow's colour indicates how far that prediction is likely to be from the true origin, on average



(a)



(b)

Fig. 7. Country classification between (a) mixed DeepSpace and (b) country Area DNN: each plot depicts the frequency of samples predicted to originate from a country (x-axis) against their true origin (y-axis)

Country Area DNN achieves a high classification rate of 84.7%. Among the three partitioning schemes, spatial models perform best with the mixed partitions which achieve country classification rates of 62.7% (Spatial NN), 74.9% (Spatial RF), 84.2% (Spatial Net) and 89.5% (DeepSpace). Fig. 7 depicts the true and predicted countries for the mixed DeepSpace and country Area DNN models. The models struggle to distinguish between samples from the bordering countries of Uruguay and Argentina, but DeepSpace misclassifies only three samples compared with 10 samples misclassified by Area DNN.

Samples for Croatia were often misclassified as being from nearby countries by DeepSpace. Interestingly, Macedonia, in contrast, which is very near to Croatia, was always correctly classified by DeepSpace. Area DNN misclassified samples from both Croatia and Macedonia, though it tended to classify them as being from nearby countries that are likely to share many fungal taxa with Croatia and Macedonia respectively. In other regions, the performance of models flipped. DeepSpace often predicted samples from Oman to be from Qatar, whereas Area DNN did not. Again, these regions are near to each other such that these errors are likely to be attributable to the shared fungal taxa among these regions. Overall, DeepSpace achieves noticeably fewer misses than does Area DNN when classifying country, despite the latter model being trained specifically for this task. This suggests that there are regional patterns within and between countries that a point level model may exploit for higher accuracy forensic geolocations.

5. Discussion

We developed a new geolocation algorithm that combines random spatial partitions with deep learning classification. The DeepSpace algorithm exploits efficient deep learning software to estimate non-parametrically a continuous Poisson intensity surface as a function of an ultrahigh dimensional predictor. DeepSpace was applied on three spatial scales: regional, national and global. The results of all geolocation methods were underwhelming in our regional analysis of the triangle region of North Carolina where, despite having over 100 samples in these three counties, none of the methods achieves over a 53% county classification accuracy. However, the DeepSpace cross-validation results are outstanding on the national and global scales, with median prediction error less than 100 km in continental USA and country classification nearly 90% for 28 countries across the globe.

A limitation of our analysis is that the sample locations were not selected randomly or systematically. The WLOH data were contributed by volunteer citizen scientists across the USA. For the global data the countries were carefully selected to represent the spatial domain of interest and the design was balanced across countries, but the sample locations within country were selected primarily on the basis of convenience. We accounted for these sampling issues in our analysis by weighting observations in the national analysis and limiting ourselves to country level predictions in the global analysis, but the sampling scheme should still be considered when interpreting and generalizing our cross-validation results. Another limitation is that we used dust samples collected from only door trims and window sills of building exteriors. These surfaces are rarely disturbed and serve as passive collectors of outdoor dust. Indoor environments and surfaces like the pavement are more regularly disturbed by human activity; therefore samples of dust from these sources may not geolocate as well as outdoor dust samples. Nonetheless, these data are more informative than any other source of data that we are aware of and the statistical contribution of the geolocation tool proposed should prove useful in other settings.

The results are for fungal data in our analysis, and a promising area of future work is to combine fungal and bacterial data. Exploratory analysis showed that fungal data are more informative than bacteria, but more sophisticated ways to incorporate bacterial data are worth

exploring. Another area of future work is to include features other than spatial co-ordinates in the predictions. For example, if there are substantial differences in the microbiomes of residential and forested areas, then predictions could be refined on the basis of the most likely land use type. Finally, we are working to include a temporal component in our analysis. A first step is to include seasonality in our geolocation algorithm, although it is likely that fungi accumulate over many months and so seasonal detection would probably require a different sampling approach (i.e. not settled dust). More ambitious is to geolocate an object that has moved in space and time. For example, the dust on a package that spent a significant amount of time in two locations may harbour a mixture of the microbiome composition at the two locations. A possible approach would be to pair the DeepSpace model with self-organizing maps (e.g. Moore *et al.* (2016)) which are widely used in forensic applications to estimate mixtures.

6. Supplementary material

Code to perform DeepSpace is written in Python 3 and is available from `https://github.com/nsgrantham/forensic-geolocation`.

Acknowledgements

This research was sponsored by the US Army Research Office and the Defense Forensic Science Center and was accomplished under co-operative agreement W911NF-16-2-0195. The views and conclusions that are contained in this paper are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, Defense Forensic Science Center or the US Government. The US Government is authorized to reproduce and distribute reprints for government purposes notwithstanding any copyright notation hereon. The work also was partially supported by National Science Foundation grants DMS-1513579 and DMS-1555141.

Appendix A: Sample collection, preparation and sequencing

Outdoor dust samples were collected with the Bode SecurSwab 2 collector or a sterile cotton swab between November 2016 and February 2017 by individuals around the world. The Bode SecurSwab 2 design enabled the collected samples to be transported back to the USA without risk that the swab head would touch the sides of the packaging which could have caused sample loss or contamination and altered the results. Dust samples were stored desiccated at room temperature and sent to the University of Colorado laboratory for molecular analysis. DNA was extracted by using the MoBio PowerSoil htp-96 Well Isolation Kit and a modified method of Barberán *et al.* (2015b). To target fungal strains, we amplified the first internal transcribed space ITS1 region of the ribosomal ribonucleic acid operon by using ITS1-F/ITS2 barcoded primers (McGuire *et al.*, 2013). Each sample was assigned a unique 12-base-pair error correcting barcode (Caporaso *et al.*, 2010) included in the ITS1 primer pairs. After polymerase chain reaction amplification in triplicate, the samples were sequenced on an Illumina MiSeq instrument running the (2×250)-base-pair MiSeq kit. Sequences were demultiplexed by using a custom Python script, pair end reads were merged by using the `usearch7 mergepairs` feature (Edgar, 2010), adapter sequences were trimmed from the merged reads by using `fastx_clipper` (`https://github.com/agordon/fastx_toolkit`) and reads were quality filtered by using `usearch7` (Edgar, 2010). Finally, sequences were clustered into operational taxonomic units by using the UPARSE pipeline (Edgar, 2013) and their taxonomic identities were determined by using a Bayesian classifier (Wang *et al.*, 2007) and compared against those in the UNITE database (Abarenkov *et al.*, 2010).

References

Abarenkov, K., Henrik Nilsson, R., Larsson, K.-H., Alexander, I. J., Eberhardt, U., Erland, S., Høiland, K., Kjøller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F. S., Tedersoo, L., Ursing, B. M., Vrålstad, T.,

Liimatainen, K., Peintner, U. and Kõljalg, U. (2010) The UNITE database for molecular identification of fungi—recent updates and future perspectives. *New Phytol.*, **186**, 281–285.
Baddeley, A., Rubak, E. and Turner, R. (2015) *Spatial Point Patterns: Methodology and Applications with R*. Boca Raton: CRC Press.
Barberán, A., Dunn, R. R., Reich, B. J., Pacifici, K., Laber, E. B., Menninger, H. L., Morton, J. M., Henley, J. B., Leff, J. W., Miller, S. L. and Fierer, N. (2015a) The ecology of microscopic life in household dust. *Proc. R. Soc.* B, **282**, article 20151139.
Barberán, A., Ladau, J., Leff, J. W., Pollard, K. S., Menninger, H. L., Dunn, R. R. and Fierer, N. (2015b) Continental-scale distributions of dust-associated bacteria and fungi. *Proc. Natn. Acad. Sci. USA*, **112**, 5756–5761.
Bryant, V. M. and Jones, G. D. (2006) Forensic palynology: current status of a rarely used technique in the United States of America. *Forens. Sci. Int.*, **163**, 183–197.
Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., Fierer, N., Peña, A. G., Goodrich, J. K., Gordon, J. I., Huttley, G. A., Kelley, S. T., Knights, D., Koenig, J. E., Ley, R. E., Lozupone, C. A., McDonald, D., Muegge, B. D., Pirrung, M., Reeder, J., Sevinsky, J. R., Turnbaugh, P. J., Walters, W. A., Widmann, J., Yatsunenko, T., Zaneveld, J. and Knight, R. (2010) Qiime allows analysis of high-throughput community sequencing data. *Nat. Meth.*, **7**, 335–336.
Chollet, F. (2015) Keras. Google AI. (Available from `https://github.com/fchollet/keras`.)
Craine, J. M., Barberán, A., Lynch, R. C., Menninger, H. L., Dunn, R. R. and Fierer, N. (2017) Molecular analysis of environmental plant DNA in house dust across the United States. *Aerobiologia*, **33**, 71–86.
Edgar, R. C. (2010) Search and clustering orders of magnitude faster than blast. *Bioinformatics*, **26**, 2460–2461.
Edgar, R. C. (2013) UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat. Meth.*, **10**, 996–998.
Friedman, J., Hastie, T. and Tibshirani, R. (2001) *The Elements of Statistical Learning*, vol. 1. New York: Springer.
Gelfand, A. E., Diggle, P., Guttorp, P. and Fuentes, M. (2010) *Handbook of Spatial Statistics*. Boca Raton: CRC Press.
Goodfellow, I., Bengio, Y. and Courville, A. (2016) *Deep Learning*. Cambridge: MIT Press.
Goodman, F., Doughty, J., Gary, C., Christou, C., Hu, B., Hultman, E., Deanto, D. and Masters, D. (2015) Piglt: a pollen identification and geolocation system for forensic applications. In *Technologies for Homeland Security (HST)*, pp. 1–7. New York: Institute of Electrical and Electronics Engineers.
Grantham, N. S., Reich, B. J., Pacifici, K., Laber, E. B., Menninger, H. L., Henley, J. B., Barberán, A., Leff, J. W., Fierer, N. and Dunn, R. R. (2015) Fungi identify the geographic origin of dust samples. *PLOS One*, **10**, no. 4, article e0122605.
Hinton, G., Deng, L., Yu, D., Dahl, G. E., Mohamed, A.-R., Jaitly, N., Senior, A., Vanhoucke, V., Nguyen, P., Sainath, T. N. and Kingsbury, B. (2012) Deep neural networks for acoustic modeling in speech recognition: the shared views of four research groups. *IEEE Signl Process. Mag.*, **29**, 82–97.
Hinton, G. E., Osindero, S. and Teh, Y.-W. (2006) A fast learning algorithm for deep belief nets. *Neurl Computn*, **18**, 1527–1554.
Jones, G. D. (2012) Forensic pollen geolocation techniques used to identify the origin of boll weevil re-infestation. *Grana*, **51**, 206–214.
Kingma, D. and Ba, J. (2014) Adam: a method for stochastic optimization. Google, Mountain View. *Preprint arXiv:1412.6980*.
LeCun, Y., Bengio, Y. and Hinton, G. (2015) Deep learning. *Nature*, **521**, 436–444.
Liang, S., Carlin, B. P. and Gelfand, A. E. (2008) Analysis of Minnesota colon and rectum cancer point patterns with spatial and non-spatial covariate information. *Ann. Appl. Statist.*, **3**, 943–962.
Locard, E. (1930) The analysis of dust traces: part I. *Am. J. Police Sci.*, **1**, 276–298.
Madden, A. A., Barberán, A., Bertone, M. A., Menninger, H. L., Dunn, R. R. and Fierer, N. (2016) The diversity of arthropods in homes across the United States as determined by environmental DNA analyses. *Molec. Ecol.*, **25**, 6214–6224.
McGuire, K. L., Payne, S. G., Palmer, M. I., Gillikin, C. M., Keefe, D., Kim, S. J., Gedallovich, S. M., Discenza, J., Rangamannar, R., Koshner, J. A., Massmann, A. L., Orazi, G., Essene, A., Leff, J. W. and Fierer, N. (2013) Digging the New York city skyline: soil fungal communities in green roofs and city parks. *PLOS One*, **8**, 1–13.
Møller, J. and Waagepetersen, R. P. (2003) *Statistical Inference and Simulation for Spatial Point Processes*. Boca Raton: CRC Press.
Moore, H. E., Butcher, J. B., Adam, C. D., Day, C. R. and Drijfhout, F. P. (2016) Age estimation of calliphora (diptera: Calliphoridae) larvae using cuticular hydrocarbon analysis and artificial neural networks. *Forens. Sci. Int.*, **268**, 81–91.
Pye, K. (2007) *Geological and Soil Evidence: Forensic Applications*. Boca Raton: CRC Press.
Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., Dieleman, S., Grewe, D., Nham, J., Kalchbrenner, N., Sutskever, I., Lillicrap, T., Leach, M., Kavukcuoglu, K., Graepel, T. and Hassabis, D. (2016) Mastering the game of Go with deep neural networks and tree search. *Nature*, **529**, 484–489.

Wang, Q., Garrity, G. M., Tiedje, J. M. and Cole, J. R. (2007) Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Micrbiol.*, **73**, 5261–5267.

Weyand, T., Kostrikov, I. and Philbin, J. (2016) Planet-photo geolocation with convolutional neural networks. In *Proc. Eur. Conf. Computer Vision* (eds B. Leibe, J. Matas, N. Sebe and M. Welling), pp. 37–55. New York: Springer.

Yue, Y. R. and Loh, J. M. (2015) Variable selection for inhomogeneous spatial point process models. *Can. J. Statist.*, **43**, 288–305.