Forensic Geolocation with Deep Neural Networks

Brian Reich
North Carolina State University
Joint work with

☞ Neal Grantham @ NCSU Statistics
☞ Eric Laber @ NCSU Statistics
☞ Kevin Gross @ NCSU Biomathematics
In collaboration with

☞ Dunn Lab @ NCSU Biological Sciences and Applied Ecology

☞ Fierer Lab @ CU Boulder Ecology & Evolutionary Biology

☞ Genomic Identification Lab @ NCSU Forensic Sciences
What is a microbiome?

Community of microbial organisms occupying an ecological niche.

Next-generation sequencing technologies make possible efficient identification of these microbes at affordable cost\(^1\).

Huge interest in understanding microbiomes as they relate to human health, diet, agriculture, environment, forensics, etc.

---

Statistical challenges

Microbiome data are...
☞ high-dimensional,
☞ sparse,
☞ over-dispersed, and
☞ possess complex dependence structure.

Many exciting opportunities for research with microbiologists.
Motivating dataset

Wild Life of Our Homes, a public science project by Rob Dunn Lab.

- Dust samples collected from outer door frames of $n \approx 1,300$ homes across the continental U.S.
- DNA sequencing revealed $p \approx 50,000$ fungal taxa.
Research question:

Does the microbiome composition of an ambient dust sample inform its geographic origin?
Our approach:

Build a model to estimate the unknown origin $s$ of a dust sample conditional on its known microbiome composition $x$. 
Initial model

Grantham et al. (2015) Fungi identify the geographic origin of dust samples. PLOS One.
Fungal diversity

k_Fungi; p_Ascomycota; c_Dothideomycetes; o_Capnidiaceae; f_Incertae sedis; g_Capnobotryella; s_Capnobotryella sp

Present?
- Yes
- No

Occurrence Probability
- 0.8
- 0.6
- 0.4
- 0.2
Geolocation

Error: 153.6 km
## Prediction error

Prediction error (km) obtained by this model:

<table>
<thead>
<tr>
<th></th>
<th>Min.</th>
<th>1st Qu.</th>
<th>Median</th>
<th>Mean</th>
<th>3rd Qu.</th>
<th>Max.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Value</td>
<td>4.268</td>
<td>149.000</td>
<td>239.300</td>
<td>403.200</td>
<td>460.600</td>
<td>4303.000</td>
</tr>
</tbody>
</table>
As seen on CSI Cyber!
New direction

Our discriminant analysis (DA) approach was mostly proof of concept.

Full DA needs the joint distribution which is too high-dimensional.

Big opportunity for more advanced statistical methods.

Here, we develop a new model based on

① spatial point pattern theory, and

② deep learning.
1. Spatial point pattern theory

Assume a spatial point pattern follows a non-homogeneous Poisson process with intensity surface $\lambda(s \mid x)$.

Liang et al.\(^6\) propose a log Gaussian process (GP),

$$\lambda(s \mid x) = \pi(s) \exp[\beta' x + w(s)]$$

with $w(s) \sim GP$, $\pi(s)$ a population offset, and $\beta$ unknown.


1. Spatial point pattern theory

Numerical approximation possible with a Monte Carlo algorithm using a knot-based predictive process but requires careful tuning.

Assumes log-linear relationship, we may need much more flexibility.

Effective in low dimensions, but no hope of fitting this model when $p \approx 50,000$.

---

2. Deep learning

Well-suited for high-dimensional data with complex structure\(^7\).

Rich, GPU-enabled software libraries available:

☞ Theano (Python)
☞ Torch7 (Lua)
☞ Tensorflow (C++)

2. Deep learning

Let $\mathcal{P} = \{P_k\}_{k=1}^K$ denote a partition of spatial domain $\mathcal{D}$.

Represent every $s$ by the region to which it belongs.

Two major benefits:

☞ Avoids costly approximation of integral in $L$.

☞ Reframes estimation as a supervised classification problem.
2. Deep learning

For regions $k = 1, \ldots, K$,

$$Pr(s \in P_k \mid x) = \frac{\exp[f_k(x)]}{\sum_{l=1}^{K} \exp[f_l(x)]} \quad \text{where } f_k : \mathcal{X} \to \mathbb{R}.$$ 

Estimate $f_1(\cdot), \ldots, f_K(\cdot)$ by training a deep neural network (DNN) on $\{(s_i, x_i)\}_{i=1}^{n}$ with categorical cross-entropy cost function

$$C = - \sum_{i=1}^{n} \sum_{k=1}^{K} \log[Pr(s_i \in P_k \mid x_i)]I(s_i \in P_k).$$
Our "Deep Space" algorithm

1. Generate random Voronoi partition $P$ over $D$.
2. Train DNN on available data from these regions.
3. Repeat steps 1 & 2 $N$ times to develop a diverse collection, $M$, of trained DNNs.
4. Predict most likely origin $\hat{s}$ by averaging over DNNs in $M$. 
1. Generate random Voronoi partition $\mathcal{P}$ over $\mathcal{D}$. 
2. Train DNN on available data from these regions.
3. Repeat steps 1 & 2 \( N \) times to develop a diverse collection, \( \mathcal{M} \), of trained DNNs.
4. Predict most likely origin $\hat{s}$ by averaging over DNNs in $\mathcal{M}$. 
Geolocation

A sample with microbiome $x$ is most likely to have originated from

$$
\hat{s} = \arg \max_{s \in \mathcal{D}} g(s \mid x, \mathcal{M})
$$

where $g(\cdot)$ is the geolocation function given by

$$
g(s \mid x, \mathcal{M}) = \sum_{j=1}^{N} w_j \sum_{k=1}^{K_j} \frac{1}{|P_{jk}|} Pr(s \in P_{jk} \mid x) I(s \in P_{jk}).
$$
Tuning

Number of seeds in each partition $K_j \sim \text{Unif}(\{30, 31, \ldots, 60\})$

Could weight DNNs by validation set performance (currently $w_j = 1/N$).

Other DNN features:

☞ Number of hidden layers (currently 4)
☞ Neurons per layer (currently 1,024)
☞ Percentage of dropout to prevent overfitting (currently 20%)
More test sample predictions...
National dust analysis

\[ n = 1,301 \] samples collected from 48 states.

\[ p = 57,331 \] unique fungal taxa.
<table>
<thead>
<tr>
<th>Model</th>
<th>10%</th>
<th>50%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSpace</td>
<td>16.8</td>
<td>86.9</td>
<td>500.4</td>
</tr>
<tr>
<td>Spatial KNN</td>
<td>22.7</td>
<td>231.9</td>
<td>953.6</td>
</tr>
<tr>
<td>Spatial RF</td>
<td>21.3</td>
<td>194.8</td>
<td>901.1</td>
</tr>
<tr>
<td>State DNN</td>
<td>58.0</td>
<td>217.2</td>
<td>606.5</td>
</tr>
</tbody>
</table>
## Area match (%)

<table>
<thead>
<tr>
<th>Model</th>
<th>State</th>
<th>County</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSpace</td>
<td>61.0</td>
<td>21.5</td>
<td>17.2</td>
</tr>
<tr>
<td>Spatial KNN</td>
<td>44.9</td>
<td>12.6</td>
<td>10.4</td>
</tr>
<tr>
<td>Spatial RF</td>
<td>48.3</td>
<td>13.9</td>
<td>11.9</td>
</tr>
<tr>
<td>State DNN</td>
<td>56.7</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>
Local dust analysis

\[ n = 116 \] samples collected from Wake, Durham, and Orange County.

\[ p = 20,557 \] unique fungal taxa.
Points with true origin captured by Mixed DS 50% pred. region
55 points of 116 total (47.4%)

Points with true origin captured by Mixed DS 75% pred. region, but not 50%
38 points of 116 total (32.6%)

Points with true origin captured by Mixed DS 90% pred. region, but not 75%
19 points of 116 total (16.4%)

Points with true origin not captured by Mixed DS 90% pred. region
12 points of 116 total (10.3%)
# Prediction errors (km)

<table>
<thead>
<tr>
<th>Model</th>
<th>10%</th>
<th>50%</th>
<th>90%</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSpace</td>
<td>7.1</td>
<td>20.0</td>
<td>41.9</td>
</tr>
<tr>
<td>Spatial KNN</td>
<td>6.4</td>
<td>20.4</td>
<td>37.0</td>
</tr>
<tr>
<td>Spatial RF</td>
<td>8.2</td>
<td>20.2</td>
<td>37.8</td>
</tr>
<tr>
<td>County DNN</td>
<td>6.7</td>
<td>21.2</td>
<td>46.6</td>
</tr>
</tbody>
</table>
### Area match (%)

<table>
<thead>
<tr>
<th>Model</th>
<th>County</th>
<th>City</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSpace</td>
<td>40.5</td>
<td>18.1</td>
</tr>
<tr>
<td>Spatial KNN</td>
<td>43.1</td>
<td>24.1</td>
</tr>
<tr>
<td>Spatial RF</td>
<td>36.2</td>
<td>19.0</td>
</tr>
<tr>
<td>County DNN</td>
<td>48.3</td>
<td>-</td>
</tr>
</tbody>
</table>
Global dust analysis

\[ n = 399 \text{ samples collected from 28 countries across 6 continents.} \]

\[ p = 15,475 \text{ unique fungal taxa.} \]
**Americas**: Mexico, Colombia, Trinidad and Tobago, Uruguay, Argeintina

**Africa**: Ghana, Nigeria, South Africa

**East Europe**: Czechia, Croatia, Hungary, Macedonia

**West Asia**: Turkey, Cyprus, Jordan

**Middle East**: Kuwait, Qatar, Oman, Georgia, Azerbaijan

**Central Asia**: Kazakhstan, Pakistan

**East Asia**: Vitenam, South Korea, Malaysia

**Oceania**: Australia, New Zealand
## Country classification (%)

<table>
<thead>
<tr>
<th>Model</th>
<th>Country</th>
</tr>
</thead>
<tbody>
<tr>
<td>DeepSpace</td>
<td>90.2</td>
</tr>
<tr>
<td>Spatial KNN</td>
<td>62.7</td>
</tr>
<tr>
<td>Spatial RF</td>
<td>74.9</td>
</tr>
<tr>
<td>Country DNN</td>
<td>84.7</td>
</tr>
</tbody>
</table>
Future work

Develop standalone software for the use of DeepSpace in practice by the U.S. Department of Defense.

Move from observational data to designed experiments to better understand:

① How dust-associated microorganisms relate to their environment, and

② How they can best be used in forensic analyses.