Testing the Correlations between Beta Diversity and Productivity using WWF WildFinder and NASA MODIS Global NDVI Datasets

Jianting Zhang

Dept of Comp. Sci. & NOAA-CREST
The City College of New York
Introduction

• Exploring the relationship between species distribution and environments is central to basic ecological research and biodiversity conservation practices.

• Considerable amount of research on the relationships between species diversity and productivity at different spatial, ecological, and taxonomic scales has been conducted. However, the overall trend of the correlation at the global scale still remains sketchy.

• Satellite products provide spatially and temporally continuous coverage at large scales which is ideal for global scale studies. NDVI has been extensively used as a surrogate for primary productivity which motivates us to use MODIS NDVI data products for testing the correlation.
Introduction

Community – Ecosystem – Biome – Biosphere

Species
Taxonomic (Linnaean ranks)
  Kingdom
  Phylum
  Class
  Order
  Family
  Genus
  Species
  SubSpecies

Phylogenetic

Environment

Area

Environmental Gradient

Water-Energy

Latitude

Altitude

Productivity

Phylogeography
Introduction

Observed/Predicted species in the regions

Regional Species Richness (Alpha Diversity)

Ecoregions: Zonal Statistics (Spatial Summation)

Regression: $Y \sim (e_1, e_2, \ldots, e_n)$

Taxonomic Dissimilarity Matrix (Beta Diversity)

Environmental Dissimilarity Matrix

Matrix Correlation Tests
Datasets

- **WWF Terrestrial Ecoregions**
  - 825 ecoregions
  - 8 biogeographical realms and 14 biomes
  - ~50M

- **WWF WildFinder Database**
  - 29,112 species, 4,815 genera, 445 families and 69 orders
  - 4 classes (amphibians, reptiles, birds, and mammals)
  - 350,045 species-ecoregion records

- **NASA NDVI**
  - Filled Normalized Difference Vegetative Index (NDVI)
  - Global data set of spatially complete NDVI maps for 23 sixteen-day periods per year
  - 2000-2004 average
  - 1 Minute (~ 2km) resolution (21600*10800)
Global Biomes and Realms

Source: [http://www.worldwildlife.org/science/data/item1875.html](http://www.worldwildlife.org/science/data/item1875.html) (Olson et al 2001)
Preprocessing of NDVI Dataset

• Data format conversion: HDF4->GeoTIFF
  – TIFF is better supported
  – Build GDAL driver for HDF4

• Retrieval of time series NDVI values for sampled pixels
  – GDAL RasterIO (C++ API)
  – Sampling strategy will be provided shortly
Pixel Sampling Strategy

WWF Ecoregion Rasterization

Count cells for each ecoregion

Choose Ecoregions with cells above 100

Random sampling and determine their row/col pairs

Generate candidate samples and their NDVI values (23 dimensions)

Remove ecoregions (and their pixels) whose sample number is below 50

Pixels and their feature vector (23D)

23 NDVI Datasets
Method

NDVI Time Series → Sampling Strategy → NDVI Samples by zones (23D) → Pair-wise Distance Matrix (Bhattacharyya)

Ecoregion → Species distribution tables

Species by zone at taxon levels → Pair-wise dissimilarity Matrix (Bray–Curtis)

Join tables and group by zones

Mantel Test

Family, Genus, Species
Method

• Distances

\[
X = \begin{bmatrix}
X_1 \\
\vdots \\
X_n
\end{bmatrix}
\]

\[
\Sigma_{ij} = \text{E}[(X_i - \mu_i)(X_j - \mu_j)]
\]

\[
\Sigma = \begin{bmatrix}
\text{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \text{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \text{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\
\text{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \text{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \text{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\
\vdots & \vdots & \ddots & \vdots \\
\text{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \text{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \text{E}[(X_n - \mu_n)(X_n - \mu_n)]
\end{bmatrix}
\]

Euclidean Distance

\[
dE = \sqrt{(x_i - x_j)^T (x_i - x_j)} \sqrt{\sum_{k=1}^{n} (x_{ik} - X_{jk})^2}
\]

Mahalanobis Distance

\[
dM = \sqrt{(x_i - x_j)^T \Sigma^{-1} (x_i - x_j)}
\]

Method

Bhattacharyya Distance (Bhatt):
- Measures dissimilarity between two groups of samples instead of two samples
- Takes both mean and variance into consideration

\[ dB = \frac{1}{8} (\mu_i - \mu_j) \left( \frac{\sum_i + \sum_j}{2} \right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} \ln \left( \frac{\sum_i + \sum_j}{2} \right) \left( \frac{\sum_i}{\sum_i^{1/2}} \frac{\sum_j}{\sum_j^{1/2}} \right) \]

Bray–Curtis Dissimilarity

\[ \beta_{BC} = \frac{\sum_{i=1}^{n} |x_{ai} - x_{bi}|}{\sum_{i=1}^{n} (x_{ai} + x_{bi})} \]

For presence/absence data: \((b+c)/(2*a+b+c)\), complementary measurement of the Sørensen similarity

Method

Mantel Test of Matrix Correlation

\[ z = \sum_{i=1}^{n} \sum_{j=1}^{n} d_{1ij} d_{2ij} \]

\[ r = \frac{1}{n-1} \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{(d_{1ij} - \bar{d}_1)(d_{2ij} - \bar{d}_2)}{S_1 S_2} \]

Permutation Test

- Permute all independent pairs (possibly through sampling)
  \[ 20! = 2432902008176640000 \]
- Using the rank of the test score among the scores of all independent pairs as the significance level
  - If a \( z/r \) value ranks \#50 among 10000 tests, then the confidence level is \( p = 50/10000 = 0.005 \)
- Vegan package in R
Results – Global

<table>
<thead>
<tr>
<th>NDVI-Species</th>
<th>Cor.</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>NDVI-Species</td>
<td>0.12872</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>NDVI-Genus</td>
<td>0.12839</td>
<td>0.0001</td>
</tr>
<tr>
<td>NDVI-Family</td>
<td>0.06797</td>
<td>&lt;0.00001</td>
</tr>
</tbody>
</table>
## Results – Global Realms

<table>
<thead>
<tr>
<th>Realm</th>
<th>No. of Ecoregions</th>
<th>NDVI-Family</th>
<th>NDVI-Genus</th>
<th>NDVI-Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cor.</td>
<td>Sig.</td>
<td>Cor.</td>
</tr>
<tr>
<td>AA</td>
<td>80</td>
<td>0.02119</td>
<td>0.3294</td>
<td>0.24165</td>
</tr>
<tr>
<td>AT</td>
<td>105</td>
<td>0.11857</td>
<td>0.0132</td>
<td>0.20153</td>
</tr>
<tr>
<td>IM</td>
<td>102</td>
<td>0.20775</td>
<td>0.0030</td>
<td>0.36345</td>
</tr>
<tr>
<td>NA</td>
<td>108</td>
<td>0.33208</td>
<td>&lt;0.0001</td>
<td>0.33479</td>
</tr>
<tr>
<td>NT</td>
<td>168</td>
<td>0.20295</td>
<td>0.0001</td>
<td>0.24989</td>
</tr>
<tr>
<td>OC</td>
<td>12</td>
<td>0.45397</td>
<td>0.0025</td>
<td>0.41631</td>
</tr>
<tr>
<td>PA</td>
<td>187</td>
<td>0.27670</td>
<td>&lt;0.0001</td>
<td>0.32669</td>
</tr>
</tbody>
</table>
# Results – Global Biomes

<table>
<thead>
<tr>
<th>Biome</th>
<th>No. of Ecoregions</th>
<th>NDVI-Family</th>
<th>NDVI-Genus</th>
<th>NDVI-Species</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Cor.</td>
<td>Sig.</td>
<td>Cor.</td>
</tr>
<tr>
<td>01</td>
<td>212</td>
<td>0.01559</td>
<td>0.2779</td>
<td>0.09360</td>
</tr>
<tr>
<td>02</td>
<td>53</td>
<td>0.08033</td>
<td>0.0840</td>
<td>0.16257</td>
</tr>
<tr>
<td>03</td>
<td>16</td>
<td>0.26428</td>
<td>0.0697</td>
<td>0.27841</td>
</tr>
<tr>
<td>04</td>
<td>82</td>
<td>0.16130</td>
<td>0.0097</td>
<td>0.18296</td>
</tr>
<tr>
<td>05</td>
<td>53</td>
<td>0.25642</td>
<td>0.0001</td>
<td>0.25675</td>
</tr>
<tr>
<td>06</td>
<td>28</td>
<td>0.23048</td>
<td>0.0237</td>
<td>0.22027</td>
</tr>
<tr>
<td>07</td>
<td>43</td>
<td>-0.03457</td>
<td>0.6420</td>
<td>0.01021</td>
</tr>
<tr>
<td>08</td>
<td>41</td>
<td>0.31976</td>
<td>&lt;0.0001</td>
<td>0.32009</td>
</tr>
<tr>
<td>09</td>
<td>21</td>
<td>0.33059</td>
<td>0.0001</td>
<td>0.32888</td>
</tr>
<tr>
<td>10</td>
<td>50</td>
<td>0.15550</td>
<td>0.0088</td>
<td>0.22306</td>
</tr>
<tr>
<td>11</td>
<td>14</td>
<td>0.72931</td>
<td>&lt;0.0001</td>
<td>0.62436</td>
</tr>
<tr>
<td>12</td>
<td>39</td>
<td>0.23208</td>
<td>0.0004</td>
<td>0.22302</td>
</tr>
<tr>
<td>13</td>
<td>92</td>
<td>0.33580</td>
<td>&lt;0.0001</td>
<td>0.32663</td>
</tr>
<tr>
<td>14</td>
<td>19</td>
<td>0.00329</td>
<td>0.4444</td>
<td>0.03544</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>0.03003</td>
</tr>
</tbody>
</table>
Summary

• This presentation focuses on:
  – Data processing: practical skills to integrate multi-source datasets and handle large-volume datasets
  – Reporting preliminary results
• Further work to interpret the results and validate the methodology is needed:
  – Work closely with ecologists, biogeographers and statisticians.
  – Identifying regions that have high positive or high negative correlations – automatic detection.
Ongoing Works

• Rasterizing Species Range Maps from NatureServe
  – Mammal (1693), Birds (4148), Amphibian (5816)
  – 1 minute spatial resolution, (360*60)*(180*60)=23,328,000 cells for each of the 11,657 species

• Designing new tree data structures for efficient representations + storing the equivalent linear data structures in PostgreSQL=> efficient query processing at the sub-polygon level

• Planned ecological/biogeographical research
  – Dissimilarity matrix correlations at the finer and multiple scales
  – Spatial correlations for both environmental and species data
  – How different are WWF Wildfinder and NatureServe data?
  – Are current ecoregion systems function as expected?
  – …..
Q&A

jzhang@cs.ccny.cuny.edu