The Contribution of Small Collections: A Case Study from Fuireneae (Cyperaceae)

Heather E. Dame\textsuperscript{1,2*}, Benjamin W. Heumann\textsuperscript{3,3}, J. Richard Carter\textsuperscript{4}, Jessica M. Bartek\textsuperscript{4}, Anna K. Monfils\textsuperscript{1,2}

(1) Central Michigan University, Institute for Great Lakes Research, Mount Pleasant, MI
(2) Central Michigan University, Department of Biology, Mount Pleasant, MI
(3) Central Michigan University, Department of Geography, Center for Geographic Information Science, Mount Pleasant, MI
(4) Department of Biology, Valdosta State University, Valdosta, GA
Uses of primary biodiversity data

- Spread of pests and invasives
- Rare species studies
- Species distribution modeling
- Historical distributions
- Climate change impacts
- Spread of disease
- Community distributions
- Food security
- Determine areas for conservation
- Phylogeography

Introduction ~ Methods ~ Results ~ Conclusions
Uses of primary biodiversity data

Species distribution modeling

- Spread of pests and invasives
- Rare species studies
- Historical distributions
- Climate change impacts
- Spread of disease
- Phylogeography
- Community distributions
- Food security
- Determine areas for conservation

Introduction ~ Methods ~ Results ~ Conclusions
Species Distribution Modeling in brief

- Allows understanding of distributions without having complete sampling of species

- Models are largely reliant on the data that is put into them
Global Biodiversity Information Facility (GBIF)

- Free, online portal for species occurrence records linked to primary biodiversity data

- Largest biodiversity database available:
  - >500 million records
  - >1.5 million species
  - Contains over 300 years of data collections
  - Cited in >1,300 peer-reviewed research publications

Vision: "A world in which biodiversity information is freely and universally available for science, society and a sustainable future."
Digitization of Small Collections

Growing appreciation for the potential contribution of small collections in the national digitization effort
Herbaria specimens in the United States

Percent specimens in size class

- **< 100,000**
  - 13%

- **≥ 100,000**
  - 87%

Small Collections

- < 100,000 specimens
- Regional collecting

Introduction ~ Methods ~ Results ~ Conclusions
Herbaria specimens in the United States

Percent specimens in size class

- < 100,000: 13%
- ≥ 100,000: 87%

Percent of herbaria in size class

- < 100,000: 83%
- ≥ 100,000: 17%
Research Question

What is the relative contribution of small collections to our understanding of species distribution and niche modeling?
Objectives

• Assess the predictive power of large, small, and combined collection datasets

• Evaluate the relative influence of large, small, and combined collection datasets on geographic predictions
Species Distribution Modeling using Maximum Entropy

MaxEnt (Phillips et al., 2006)

- Predict suitable habitat over a geographic space

- Presence-only modeling method

- Consistent high performance among other modeling methods
Modeling extent of the Contiguous U.S.

Consistently high quality environmental variables
MaxEnt builds a landscape showing the probability of suitable habitat

- Creates background data grid
- Correlates presence points to background grid cells
- Defines mean and variance of niche
- Builds predictions of the probability of suitable habitat
Methodology

• Obtain Species Occurrence Data
• Filter Data
• Species Distribution Model
• Model Prediction Evaluation
• Geographic Space Analysis
Methodology

• Obtain Species Occurrence Data
• Filter Data
• Species Distribution Model
• Model Prediction Evaluation
• Geographic Space Analysis
Two sources of collections data

• GBIF collections data
  • Routinely used in species distributions modeling studies

• Small regional collections collaborations
  • Central Michigan University (CMC) and Valdosta State University (VSC)
Fuiereae (Cyperaceae; Sedges)

• Wetland plants

• 4 genera naturally occur in the United States
  • Wide ranging
  • Narrow endemics
Methodology

- Obtain Species Occurrence Data
- Filter Data
- Species Distribution Model
- Geographic Space Analysis
- Model Prediction Evaluation
Three Independent Datasets

- GBIF Large Collections
  - Occurrences from collections >100,000

- GBIF Small Collections
  - Occurrences from collections <100,000

- CMC/VSC Collections
  - Occurrences from CMC and VSC herbaria
Data quality

- Retained data with sufficient metadata
  - Preserved Specimens
  - No voucher duplicates
  - Georeferenced quality
    - Georeferenced CMC / VSC collections using GeoLocate (Rios & Bart, 2010)
Selecting species

• ≥ 10 occurrence records
• Species present in both GBIF and CMC/VSC datasets
• No obligate halophytes
• Removed geographic replicates using ENMTools (Warren et al., 2010)
## Species selected for analysis

<table>
<thead>
<tr>
<th>Species</th>
<th>GBIF Large</th>
<th>GBIF Small</th>
<th>CMC/VSC</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Fuirena pumila</em> (Torrey) Sprengel</td>
<td>10</td>
<td>n/a</td>
<td>22</td>
<td>32</td>
</tr>
<tr>
<td><em>Fuirena squarrosa</em> Michx.</td>
<td>25</td>
<td>n/a</td>
<td>44</td>
<td>69</td>
</tr>
<tr>
<td><em>Schoenoplectiella purshiana</em> (Fernald) Lye</td>
<td>45</td>
<td>n/a</td>
<td>15</td>
<td>60</td>
</tr>
<tr>
<td><em>Schoenoplectus acutus</em> (Bigelow) Á. Löve &amp; D. Löve</td>
<td>434</td>
<td>52</td>
<td>13</td>
<td>499</td>
</tr>
<tr>
<td><em>Schoenoplectus pungens</em> (Vahl) Palla</td>
<td>413</td>
<td>32</td>
<td>26</td>
<td>471</td>
</tr>
<tr>
<td><em>Schoenoplectus tabernaemontani</em> (C.C. Gmel.) Palla</td>
<td>352</td>
<td>38</td>
<td>29</td>
<td>419</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>1270</strong></td>
<td><strong>122</strong></td>
<td><strong>149</strong></td>
<td><strong>1550</strong></td>
</tr>
</tbody>
</table>

**Introduction ~ Methods ~ Results ~ Conclusions**
**Fuirena pumila**
Dwarf umbrella sedge

**Schoenoplectus acutus**
Hardstem bulrush

**Fuirena squarrosa**
Hairy umbrella sedge

**Schoenoplectus pungens**
Common Threesquare

**Schoenoplectiella purshiana**
Weakstalk bulrush

**Schoenoplectus tabernaemontani**
Soft-stem bulrush
Geographic locations of occurrences

GBIF Large Dataset Occurrences

GBIF Small Dataset Occurrences

CMC/VSC Dataset Occurrences

CMC Herbarium (Michigan)

VSC Herbarium (Georgia)
Environmental Factors

WorldClim Bioclimatic Factors (Hijmans et al. 2005)
• 7 factors

STATSGO2 Soil Factors (Soil Survey Staff)
• 7 factors

Total: 14 environmental variables
Methodology

- Obtain Species Occurrence Data
- Filter Data
- Species Distribution Model
- Model Prediction Evaluation
- Geographic Space Analysis
Three Additive Data Inputs

- Occurrences from GBIF Large collections
- Occurrences from both GBIF Small and Large collections
- Occurrences from all GBIF and CMC/VSC collections
Fuirena pumila
Dwarf umbrella sedge

Fuirena squarrosa
Hairy umbrella sedge

Schoenoplectiella purshiana
Weakstalk bulrush

Schoenoplectus acutus
Hardstem bulrush

Schoenoplectus pungens
Common Threesquare

Schoenoplectus tabernaemontani
Soft-stem bulrush

Maps: floranorthamerica.org
Photos: MichiganFlora.net and Plants.USDA.org
**Fuirena pumila**  
Dwarf umbrella sedge

**Schoenoplectus purshiana**  
Weakstalk bulrush

**Schoenoplectus acutus**  
Hardstem bulrush

**Schoenoplectus tabernaemontani**

Maps: floranorthamerica.org  
Photos: MichiganFlora.net and Plants.USDA.org
MaxEnt maps displaying the probability of suitable habitat

Schoenoplectus tabernaemontani

Model built from GBIF Large Collections

Model built from GBIF Small & Large Collections

Model built from all GBIF Collections & CMC/VSC
Methodology

- Obtain Species Occurrence Data
- Filter Data
- Species Distribution Model
- Model Prediction Evaluation
- Geographic Space Analysis
Comparing model results displayed significant differences between extracted probabilities of suitable habitat

Friedman test results

\[ p < 0.05 \]

\[ S. \text{tabernaemontani} \]
Comparing model results displayed significant differences between extracted probabilities of suitable habitat.

GBIF Small collections occurrences

Friedman test results $p < 0.005$

*S. tabernaemontani*
Comparing model results displayed significant differences between extracted probabilities of suitable habitat.

**CMC/VSC Collections**

Friedman test results

\[ p < 0.005 \]

*S. tabernaemontani*
Methodology

• Obtain Species Occurrence Data
• Filter Data
• Species Distribution Model
• Model Prediction Evaluation
• Geographic Space Analysis
MaxEnt maps displaying the probability of suitable habitat

Model built from GBIF Large Collections

Model built from GBIF Small & Large Collections

Model built from all GBIF Collections & CMC/VSC

Schoenoplectus tabernaemontani

Probability of Suitable Habitat

- 0.0 - 0.2
- 0.2 - 0.4
- 0.4 - 0.6
- 0.6 - 0.8
- 0.8 - 1

Introduction ~ Methods ~ Results ~ Conclusions
Comparing geographic predictions between each model
Differences when occurrences from GBIF small collections are added to GBIF large collection based models

[Schoenoplectus tabernaemontani]

Introduction ~ Methods ~ Results ~ Conclusions
Comparing geographic predictions between each model
Differences when occurrences from all small collections are added to large collection based models.
Comparing geographic predictions between each model
Differences when occurrences from CMC/VSC collections are added to GBIF large and small collections based models.
Comparison of models predicted probability based on different datasets

GBIF Large relative to GBIF Large and Small collections

- 29%
- 13%
- 58%

GBIF Large relative to GBIF Large, Small, and CMC/VSC collections

- 22%
- 23%
- 55%

GBIF Large and Small relative to GBIF Large, Small, and CMC/VSC collections

- 28%
- 13%
- 59%

0 - 2% Difference  2 – 10% Difference  10 – 50% Difference
Small collections are strong contributors to models of species distribution and niche models

• Models inclusive of small collections data resulted in statistically significant increases in occurrence predictions

• Models inclusive of small collections data resulted in a 23% major (10-50%) change in geographic predictions
Small is Big!

• Small collections significantly refine species distribution models

• These collections may represent a small 13% of national specimens, but they are critical to building our understanding of habitats and biodiversity

Remember: there are no small parts, only small actors

— Constantin Stanislavski
Questions?
Accounting for sampling bias

- Bias file
  - Quantity of sampling across background
  - Samples background data from weighted cells
Niche comparison

• Extracted environmental data at each occurrence point for 3 independent datasets

• Compared each variable’s set of values among datasets using the Kolmogorov-Smirnov test.
  • $p < 0.05$ = a single dataset alone does not contain the “true” realized niche of a species.
Results: Niche comparison

Number of significant variable differences between all datasets

Dataset comparisons
- Large vs. CMC/VSC
- Large vs Small
- Small vs. CMC/VSC

Species
1. Fuirena pumila
2. Fuirena squarrosa
3. Schoenoplectella purshiana
4. Schoenoplectus acutus
5. Schoenolecutus pungens
6. Schoenoplectus tabernasianus

Number of variables $p < 0.05$
Filtering occurrence records

GBIF records

44 taxa
15,967 records

CMC / VSC records

26 taxa
538 records

Fuireneae taxa

Use records based on "Preserved Specimen"

Remove coarse GPS resolution records

Constrain data to contiguous United States

Reconcile taxonomy of all taxa

Remove taxa ID’d as hybrid or genus only

Remove duplicates

Remove species with <10 records

Keep species present in GBIF and CMC/VSC

Remove obligate halophytes

Remove grid duplicates by dataset

Final Dataset

GBIF Large

1279 records

GBIF Small

122 records

CMC / VSC

149 records

6 Species
### Significant differences between predicted distributions are present

<table>
<thead>
<tr>
<th>Species</th>
<th>Occurrence</th>
<th>Wilcoxon P-value</th>
<th>Friedman P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Fuirena pumila</em></td>
<td>Large</td>
<td>&lt; 0.05</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CMC/VSC</td>
<td>&lt; 0.005</td>
<td>-</td>
</tr>
<tr>
<td><em>Fuirena squarrosa</em></td>
<td>Large</td>
<td>&gt; 0.05</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CMC/VSC</td>
<td>&lt; 0.005</td>
<td>-</td>
</tr>
<tr>
<td><em>Schoenoplectiella purshiana</em></td>
<td>Large</td>
<td>&lt; 0.005</td>
<td>-</td>
</tr>
<tr>
<td></td>
<td>CMC/VSC</td>
<td>&lt; 0.005</td>
<td>-</td>
</tr>
<tr>
<td><em>Schoenoplectus acutus</em></td>
<td>Large</td>
<td>-</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>-</td>
<td>&gt; 0.05</td>
</tr>
<tr>
<td></td>
<td>CMC/VSC</td>
<td>-</td>
<td>&gt; 0.05</td>
</tr>
<tr>
<td><em>Schoenoplectus pungens</em></td>
<td>Large</td>
<td>-</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>-</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>CMC/VSC</td>
<td>-</td>
<td>&lt; 0.005</td>
</tr>
<tr>
<td><em>Schoenoplectus tabernaemontani</em></td>
<td>Large</td>
<td>-</td>
<td>&lt; 0.05</td>
</tr>
<tr>
<td></td>
<td>Small</td>
<td>-</td>
<td>&lt; 0.005</td>
</tr>
<tr>
<td></td>
<td>CMC/VSC</td>
<td>-</td>
<td>&lt; 0.005</td>
</tr>
</tbody>
</table>
Number of Herbaria in each size class

- 10K-49K: 37%
- 200-9K: 36%
- > 500K: 5%
- 100K - 490K: 12%
- 50K - 99K: 10%
Species Distribution Modeling

Introduction ~ Methods ~ Results ~ Conclusions
Assessing likelihood of suitable habitat by dataset

- GBIF Large Model
- GBIF All Model
- All GBIF & CMC/VSC Model

Extract suitability by independent dataset

Test differences within datasets

- No difference in distributions
- Difference in distributions

- Wilcoxon test
- Friedman test

Introduction ~ Methods ~ Results ~ Conclusions
Species Distribution Modeling in brief

• Allows understanding of distributions without having complete sampling of species

• Modeling studies:
  • Habitat suitability modeling (ex. Abdi, 2013; Ballesteros-Mejia et al. 2013)
  • Historical speciation patterns (ex. Liu et al. 2013)
  • Invasive species potential distributions (ex. Gallardo et al 2013)
  • Environmental variable impacts (ex. Ortega & Ober, 2013)
  • Distributions under climate change (ex. Jueterbock et al., 2013; Kriticos et al., 2013)

• Models are reliant on the data that is put into them.
Species Distribution Modeling

- Species Occurrences & Environmental Variables
- MaxEnt Model
- Geographic Output
- Geographic Space Analysis
- Model Prediction Evaluation

Introduction ~ Methods ~ Results ~ Conclusions
References
