Eliciting and encoding expert knowledge on variable selection into species distribution models (SDMs)

R. Pirathiban\textsuperscript{1}  K. J. Williams\textsuperscript{2}  A. N. Pettitt\textsuperscript{1}  S. J. Low Choy\textsuperscript{13}

\textsuperscript{1}School of Mathematical Sciences
Queensland University of Technology

\textsuperscript{2}CSIRO Land and Water
Canberra

\textsuperscript{3}Griffith Social and Behavioural Research College
Griffith University

The International Biometric Society Australasian Region Conference
29 Nov - 3 Dec, 2015
Outline

1 Introduction
   • Background
   • Current approaches for variable selection in SDMs
   • Bayesian SDMs and variable selection methods
   • Aim

2 Case study

3 Method
   • Elicitation strategy
   • Encoding model

4 Results

5 Conclusion and discussion
Quality of SDMs relies on the quality of the input data, from bioclimatic indices to environmental and habitat descriptors.
Current approaches for variable selection in SDMs

A priori selection of variables

- Environmental niche models  
  Nix (1986)

- Generalized linear model without variable selection  
  Miller & Franklin (2002)

Explicit variable selection

- Generalized linear/additive models with variable selection  
  Hastie et al. (2002)

- Classification trees with complexity/model-based pruning  
  Breiman et al. (1984), Zeileis et al. (2008)

Model averaging

- Neural networks  
  Stockwell (1999)

- Boosted/ bagged regression trees  
  Leathwick et al. (2006)

- Maximum Entropy  
  Phillips et al. (2006)

Researchers either consider the first approach with some variables or the second or third approaches with all the candidate variables
Limitations

- Does not necessarily select the best set of explanatory variables
- Investigating all possible combinations of variables is complex (e.g. 5 variables $\rightarrow 2^5 = 32$, 10 variables $\rightarrow 2^{10} = 1024$)
- Known tendency for under-fitting/over-fitting

Solution

Incorporating expert knowledge into variable selection
Elicitation approaches in Bayesian SDMs

Bayesian framework provides explicit mechanism to include expert knowledge through priors

Bayesian SDMs

- Classification trees (O’Leary et al. 2008)
- Hierarchical models (e.g. conditional probability networks) (Marcot et al. 2006, McCann et al. 2006)

Focused on

Elicitation of model parameters/one model structure NOT variable importance

One exception

Bayesian classification and regression trees (CART) (O’Leary et al. 2008)
Bayesian variable selection in Regression models

- **Spike and slab**
  (Mitchell & Beauchamp 1988)

- **Laplace**
  (Frühwirth-Schnatter & Wagner 2011)

- **Lasso models**
  (Park & Casella 2008)

---

**Indicator variable selection models**
(Kuo & Mallick 1998)

**Ridge regression**
Aim

To facilitate variable selection in species distribution models via Bayesian informative priors, constructed from the knowledge elicited from experts

- Construct an elicitation protocol that can extract the knowledge from experts
- Focus on ways to restructure the priors to encode elicited information
Eucalyptus tenuiramis

- Commonly known as silver peppermint
- Endemic species, locally common in south-eastern and eastern Tasmania
- 1442 presences and 7165 absences
- 31 environmental covariates which is a mixture of climatic (5), topographic (1) and soil (25) variables

Source: Williams & Potts (1996)
Elicitation strategy

- Developed incorporating the six main features of elicitation
  (Low Choy et al. 2009)
- Univariate and Absolute elicitation of the importance of variables
  (O’Leary et al. 2008)

Ranking

A simple ordering of variables from optimum to the worst

*Let’s sort all the soil variables according to the importance of deciding the habitat suitability of* Eucalyptus tenuiramis *from the most significant to the least significant*
Encoding model

Model 1:

Indicator variable selection model - Independent Bernoulli-Beta prior

\[ Y_i \sim \text{Bern}(\mu_i) \]

\[ \text{logit}(\mu_i) = \beta_0 + \sum_{j\in j_0} \delta_{ij} \beta_{ij} X_{ij} + \sum_{j\in j_1} \beta_{ij} X_{ij} \]

\[ \delta_{ij} \sim \text{Bern}(p_j) \]

\[ p_j \sim \text{Beta}(1, 1) \]

\[ \beta_0, \beta_{ij}, \beta_{ik} \sim \text{N}(0, 1000) \]
Model 2:

Indicator variable selection model - Ranks encoded as inclusion probability on Bernoulli prior

\[ Y_i \sim \text{Bern}(\mu_i) \]

\[
\text{logit}(\mu_i) = \beta_0 + \sum_{j \in j_0} \delta_{ij} \beta_{ij} X_{ij} + \sum_{j \in j_1} \beta_{ij} X_{ij}
\]

\[
\delta_{ij} \sim \text{Bern}(p_j)
\]

\[
\beta_0, \beta_{ij}, \beta_{ik} \sim N(0, 1000)
\]
## Elicited variable importance

<table>
<thead>
<tr>
<th>Variable</th>
<th>Description</th>
<th>Ranks</th>
</tr>
</thead>
<tbody>
<tr>
<td>geollmnage</td>
<td>Mean in Log10 geological age</td>
<td>10</td>
</tr>
<tr>
<td>geollrnage</td>
<td>Range log10 geological age</td>
<td>10</td>
</tr>
<tr>
<td>gravity9se</td>
<td>Bouger gravity anamalies</td>
<td>11</td>
</tr>
<tr>
<td>magnetic9s</td>
<td>Magnetic anomalies</td>
<td>11</td>
</tr>
<tr>
<td>nutrientsn</td>
<td>Nutrient status</td>
<td>9</td>
</tr>
<tr>
<td>minfertfe</td>
<td>Lithology - inherent fertility rating</td>
<td>9</td>
</tr>
<tr>
<td>pawc1me</td>
<td>Soils - plant available water holding capacity</td>
<td>1</td>
</tr>
<tr>
<td>ill20ne</td>
<td>Illite clay minerals in surficial topsoil</td>
<td>4</td>
</tr>
<tr>
<td>ill80ne</td>
<td>Illite clay minerals in surficial subsoil</td>
<td>4</td>
</tr>
<tr>
<td>kao20ne</td>
<td>Kaolinite clay minerals in surficial topsoil</td>
<td>6</td>
</tr>
<tr>
<td>kao80ne</td>
<td>Kaolinite clay minerals in surficial subsoil</td>
<td>6</td>
</tr>
<tr>
<td>sme20ne</td>
<td>Smectite clay minerals in surficial topsoil</td>
<td>6</td>
</tr>
<tr>
<td>sme80ne</td>
<td>Smectite clay minerals in surficial subsoil</td>
<td>6</td>
</tr>
<tr>
<td>pc1_20ne</td>
<td>Spectra of surficial topsoils–Principal component 1</td>
<td>5</td>
</tr>
<tr>
<td>pc1_80ne</td>
<td>Spectra of surficial subsoils–Principal component 1</td>
<td>5</td>
</tr>
<tr>
<td>pc2_20ne</td>
<td>Spectra of surficial topsoils–Principal component 2</td>
<td>7</td>
</tr>
<tr>
<td>pc2_80ne</td>
<td>Spectra of surficial subsoils–Principal component 2</td>
<td>7</td>
</tr>
<tr>
<td>pc3_20ne</td>
<td>Spectra of surficial topsoils–Principal component 3</td>
<td>8</td>
</tr>
<tr>
<td>pc3_80ne</td>
<td>Spectra of surficial subsoils–Principal component 3</td>
<td>8</td>
</tr>
<tr>
<td>ksatne</td>
<td>Hydrologic conductivity</td>
<td>14</td>
</tr>
<tr>
<td>bd30e</td>
<td>Soils - bulk density</td>
<td>2</td>
</tr>
<tr>
<td>hstructne</td>
<td>Pedality hydrological score</td>
<td>13</td>
</tr>
<tr>
<td>soldepthn</td>
<td>Solum depth</td>
<td>3</td>
</tr>
<tr>
<td>clay30e</td>
<td>Soils - clay fraction</td>
<td>1</td>
</tr>
<tr>
<td>wlioz2_w9s</td>
<td>weathering intensity index</td>
<td>12</td>
</tr>
</tbody>
</table>

- 25 soil variables ranked according to their order of importance on deciding the habitat suitability for *Eucalyptus tenuiramis*
Model 1: Non-expert informed variable selection

Figure: Soil variable subsets and their posterior probability via $p(\delta|\ldots)$

Figure: Soil variables colored based on top most model, in top 5 models, $\delta$ not significant
Model 1: Non-expert informed variable selection

Figure: Soil variable subsets and their posterior probability via $p(\delta|...)$

Figure: Soil variables colored based on top most model, in top 5 models, $\delta$ not significant
### Results

**Model2: Expert informed variable selection**

**Figure:** Soil variable subsets and their posterior probability via $p(\delta | ...)$

**Figure:** Soil variables colored based on top most model, in top 5 models
Conclusion

- Bayesian framework- explicit and formal mechanism for incorporating expert knowledge
- Indicator variable selection model- explicit means of variable selection
- Informative priors influences the variable selection model to some extend

Current work

- Extend the elicitation protocol to capture more on variable importance
- Restructure the priors to encode the elicited information
References I


Kynn, M. (2005), Eliciting expert knowledge for Bayesian logistic regression in species habitat modelling, PhD thesis, Queensland University of Technology.


