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

R. Pirathiban[1] **K. J. Williams**[2] **A. N. Pettitt**[1] **S. J. Low Choy**[1 3]

[1]School of Mathematical Sciences
Queensland University of Technology

[2]CSIRO Land and Water
Canberra

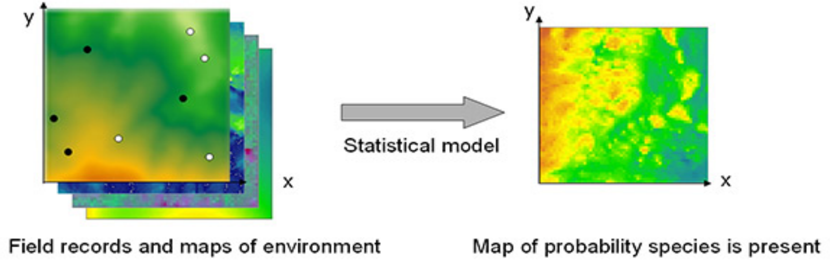
[3]Griffith Social and Behavioural Research College
Griffith University

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



Source: http://www.biodiversityscience.com/2011/04/27/species-distribution-modelling/

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 -> $2^5 = 32$, 10 variables -> $2^{10} = 1024$)
- 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

- Logistic regression models(Kynn 2005, Denham & Mengersen 2007, Murray et al. 2009)
- 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

Indicator variable selection models

(Kuo & Mallick 1998)

- Spike and slab

  (Mitchell & Beauchamp 1988)
- Laplace

  (Frühwirth-Schnatter & Wagner 2011)
- Lasso models

  (Park & Casella 2008)

Ridge regression

Aim

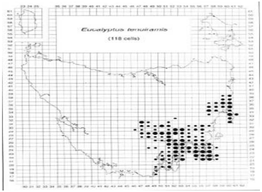
To facilitate variable selection in species distribution models via Bayesian informative priors, constructed from the knowledge elicited from experts

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

Eucalyptus tenuiramis

Figure: Adult leaves

Photo: © Greg Jordan



Source: Williams & Potts (1996)

- 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

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

Model1:

Indicator variable selection model - Independent Bernoulli-Beta prior

$$Y_i \sim Bern(\mu_i)$$

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

$$\delta_{ij} \sim Bern(p_j)$$

$$p_j \sim Beta(1, 1)$$

$$\beta_0, \beta_{ij}, \beta_{ik} \sim N(0, 1000)$$

Model2:

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

$$Y_i \sim Bern(\mu_i)$$

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

$$\delta_{ij} \sim Bern(p_j)$$

$$\beta_0, \beta_{ij}, \beta_{ik} \sim N(0, 1000)$$

Elicited variable importance

Variable	Description	Ranks
geollmnage	Mean in Log10 geological age	10
geollrnage	Range log10 geological age	10
gravity9se	Bouger gravity anamalies	11
magnetic9s	Magnetic anomalies	11
nutrientsn	Nutrient status	9
minfertfe	Lithology - inherent fertility rating	9
pawc1me	Soils - plant available water holding capacity	1
ill20ne	Illite clay minerals in surficial topsoil	4
ill80ne	Illite clay minerals in surficial subsoil	4
kao20ne	Kaolinite clay minerals in surficial topsoil	6
kao80ne	Kaolinite clay minerals in surficial subsoil	6
sme20ne	Smectite clay minerals in surficial topsoil	6
sme80ne	Smectite clay minerals in surficial subsoil	6
pc1_20ne	Spectra of surficial topsoils–Principal component 1	5
pc1_80ne	Spectra of surficial subsoils–Principal component 1	5
pc2_20ne	Spectra of surficial topsoils–Principal component 2	7
pc2_80ne	Spectra of surficial subsoils–Principal component 2	7
pc3_20ne	Spectra of surficial topsoils-Principal component 3	8
pc3_80ne	Spectra of surficial subsoils-Principal component 3	8
ksatne	Hydrologic conductivity	14
bd30e	Soils - bulk density	2
hstructne	Pedality hydrological score	13
soldepthne	Solum depth	3
clay30e	Soils - clay fraction	1
wiioz2_w9s	weathering intensity index	12

- 25 soil variables ranked according to their order of importance on deciding the habitat suitability for *Eucalyptus tenuiramis*

Model1: Non-expert informed variable selection

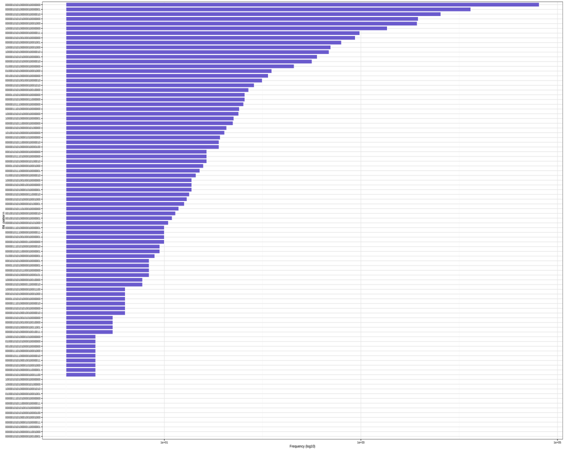


Figure: Soil variable subsets and their posterior probability via $p(\delta|...)$

Variable	Description	Variable number	Ranks
geollmnage	Mean in Log10 geological age	1	10
geollrnage	Range log10 geological age	2	10
gravity9se	Bouger gravity anamalies	3	11
magnetic9s	Magnetic anomalies	4	11
nutrientsn	Nutrient status	5	9
minfertfe	Lithology - inherent fertility rating	6	9
pawc1me	Soils - plant available water holding capacity	7	1
ill20ne	Illite clay minerals in surficial topsoil	8	4
ill80ne	Illite clay minerals in surficial subsoil	9	4
kao20ne	Kaolinite clay minerals in surficial topsoil	10	6
kao80ne	Kaolinite clay minerals in surficial subsoil	11	6
sme20ne	Smectite clay minerals in surficial topsoil	12	6
sme80ne	Smectite clay minerals in surficial subsoil	13	6
pc1_20ne	Spectra of surficial topsoils–Principal component 1	14	5
pc1_80ne	Spectra of surficial subsoils–Principal component 1	15	5
pc2_20ne	Spectra of surficial topsoils–Principal component 2	16	7
pc2_80ne	Spectra of surficial subsoils–Principal component 2	17	7
pc3_20ne	Spectra of surficial topsoils-Principal component 3	18	8
pc3_80ne	Spectra of surficial subsoils-Principal component 3	19	8
ksatne	Hydrologic conductivity	20	14
bd30e	Soils - bulk density	21	2
hstructne	Pedality hydrological score	22	13
soldepthne	Solum depth	23	3
clay30e	Soils - clay fraction	24	1
wiioz2_w9s	weathering intensity index	25	12

Figure: Soil variables colored based on top most model, in top 5 models, δ not significant

Model1: Non-expert informed variable selection

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

Variable	Description	Variable number	Ranks
geollmnage	Mean in Log10 geological age	1	10
geollrnage	Range log10 geological age	2	10
gravity9se	Bouger gravity anamalies	3	11
magnetic9s	Magnetic anomalies	4	11
nutrientsn	Nutrient status	5	9
minfertfe	Lithology - inherent fertility rating	6	9
pawc1me	Soils - plant available water holding capacity	7	1
ill20ne	Illite clay minerals in surficial topsoil	8	4
ill80ne	Illite clay minerals in surficial subsoil	9	4
kao20ne	Kaolinite clay minerals in surficial topsoil	10	6
kao80ne	Kaolinite clay minerals in surficial subsoil	11	6
sme20ne	Smectite clay minerals in surficial topsoil	12	6
sme80ne	Smectite clay minerals in surficial subsoil	13	6
pc1_20ne	Spectra of surficial topsoils–Principal component 1	14	5
pc1_80ne	Spectra of surficial subsoils–Principal component 1	15	5
pc2_20ne	Spectra of surficial topsoils–Principal component 2	16	7
pc2_80ne	Spectra of surficial subsoils–Principal component 2	17	7
pc3_20ne	Spectra of surficial topsoils-Principal component 3	18	8
pc3_80ne	Spectra of surficial subsoils-Principal component 3	19	8
ksatne	Hydrologic conductivity	20	14
bd30e	Soils - bulk density	21	2
hstructne	Pedality hydrological score	22	13
soldepthne	Solum depth	23	3
clay30e	Soils - clay fraction	24	1
wiioz2_w9s	weathering intensity index	25	12

Figure: Soil variables colored based on top most model, in top 5 models, δ not significant

Model2: Expert informed variable selection

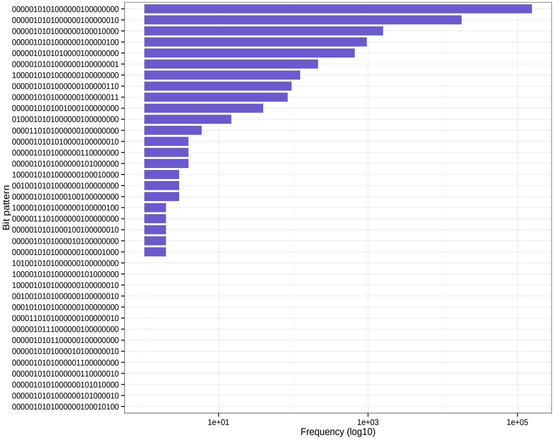


Figure: Soil variable subsets and their posterior probability via $p(\delta|\ldots)$

Variable	Description	Variable number	Ranks
geollmnage	Mean in Log10 geological age	1	10
geollrnage	Range log10 geological age	2	10
gravity9se	Bouger gravity anamalies	3	11
magnetic9s	Magnetic anomalies	4	11
nutrientsn	Nutrient status	5	9
minfertfe	Lithology - inherent fertility rating	6	9
pawc1me	Soils - plant available water holding capacity	7	1
ill20ne	Illite clay minerals in surficial topsoil	8	4
ill80ne	Illite clay minerals in surficial subsoil	9	4
kao20ne	Kaolinite clay minerals in surficial topsoil	10	6
kao80ne	Kaolinite clay minerals in surficial subsoil	11	6
sme20ne	Smectite clay minerals in surficial topsoil	12	6
sme80ne	Smectite clay minerals in surficial subsoil	13	6
pc1_20ne	Spectra of surficial topsoils–Principal component 1	14	5
pc1_80ne	Spectra of surficial subsoils–Principal component 1	15	5
pc2_20ne	Spectra of surficial topsoils–Principal component 2	16	7
pc2_80ne	Spectra of surficial subsoils–Principal component 2	17	7
pc3_20ne	Spectra of surficial topsoils-Principal component 3	18	8
pc3_80ne	Spectra of surficial subsoils-Principal component 3	19	8
ksatne	Hydrologic conductivity	20	14
bd30e	Soils - bulk density	21	2
hstructne	Pedality hydrological score	22	13
soldepthne	Solum depth	23	3
clay30e	Soils - clay fraction	24	1
wiioz2_w9s	weathering intensity index	25	12

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

Breiman, L., Friedman, J., Stone, C. J. & Olshen, R. A. (1984), *Classification and regression trees*, CRC press.

Denham, R. & Mengersen, K. (2007), 'Geographically assisted elicitation of expert opinion for regression models', *Bayesian Analysis* **2**(1), 99–136.

Frühwirth-Schnatter, S. & Wagner, H. (2011), Bayesian variable selection for random intercept modeling of gaussian and non-gaussian data, *in* J. M. Bernardo, M. Bayarri, J. O. Berger, A. Dawid, D. Heckerman & A. F. Smith, eds, 'Bayesian Statistics 9', Vol. 9, Oxford University Press, pp. 165–200.

Hastie, T., Tibshirani, R. & Friedman, J. (2002), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, second edition, fifth reprinting edn, Springer-Verlag. Available online at `http://www.stanford.edu/~hastie/local.ftp/Springer/ESLII_print5.pdf`, accessed 3 April 2012.

Kuo, L. & Mallick, B. (1998), 'Variable selection for regression models', *Sankhyā: The Indian Journal of Statistics, Series B* pp. 65–81.

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

Leathwick, J., Elith, J. & Hastie, T. (2006), 'Comparative performance of Generalized Additive Models and Multivariate Adaptive Regression Splines for statistical modelling of species distributions', *Ecological modelling* **199**(2), 188–196.

Low Choy, S., O'Leary, R. & Mengersen, K. (2009), 'Elicitation by design in ecology: Using expert opinion to inform priors for Bayesian statistical models', *Ecology* **90**(1), 265–277.

Marcot, B. G., Steventon, J. D., Sutherland, G. D. & McCann, R. K. (2006), 'Guidelines for developing and updating Bayesian Belief Networks applied to ecological modeling and conservation', *Canadian Journal of Forest Research* **36**(12), 3063–3074.

McCann, R. K., Marcot, B. G. & Ellis, R. (2006), 'Bayesian Belief Networks: applications in ecology and natural resource management', *Canadian Journal of Forest Research* **36**(12), 3053–3062.

Miller, J. & Franklin, J. (2002), 'Modeling the distribution of four vegetation alliances using Generalized Linear Models and Classification Trees with spatial dependence', *Ecological Modelling* **157**(2), 227–247.

References II

Mitchell, T. J. & Beauchamp, J. J. (1988), 'Bayesian variable selection in linear regression', *Journal of the American Statistical Association* **83**(404), 1023–1032.

Murray, J. V., Goldizen, A. W., O'Leary, R. A., McAlpine, C. A., Possingham, H. P. & Low Choy, S. (2009), 'How useful is expert opinion for predicting the distribution of a species within and beyond the region of expertise? a case study using brush-tailed rock-wallabies Petrogale penicillata', *Journal of Applied Ecology* **46**(4), 842–851.

Nix, H. A. (1986), 'A biogeographic analysis of Australian elapid snakes', *Atlas of Elapid Snakes of Australia.(Ed.) R. Longmore* pp. 4–15.

O'Leary, R. A., Murray, J. V., Low Choy, S. J. & Mengersen, K. L. (2008), 'Expert elicitation for Bayesian classification trees', *Journal of Applied Probability & Statistics* **3**(1), 95–106.

Park, T. & Casella, G. (2008), 'The Bayesian lasso', *Journal of the American Statistical Association* **103**(482), 681–686.

Phillips, S. J., Anderson, R. P. & Schapire, R. E. (2006), 'Maximum entropy modeling of species geographic distributions', *Ecological modelling* **190**(3), 231–259.

Stockwell, D. R. (1999), Genetic algorithms ii, *in* 'Machine learning methods for ecological applications', Springer, pp. 123–144.

Williams, K. & Potts, B. (1996), 'The natural distribution of eucalyptus species in tasmania', *Tasforests* **8**, 39–165.

Zeileis, A., Hothorn, T. & Hornik, K. (2008), 'Model-based recursive partitioning', *Journal of Computational and Graphical Statistics* **17**(2), 492–514.