Ecoregion Classification Using a Bayesian Approach and Model-based Clustering

D. Pullar[a] **S. Low Choy**[b], **W. Rochester**[a]

[a] ***Geography Planning and Architecture, The University of Queensland, Brisbane QLD 4072, Australia***

[b]***Environmental Information Systems, Environmental Protection Agency, PO Box 155, Albert Street, Brisbane QLD 400, Australia.*** **E-Mail .*D.Pullar@uq.edu.au***

Keywords*:* *Biogeography; Bayesian statistical modelling; GIS; Elicitation; Mixture models; Cluster*ing

EXTENDED ABSTRACT

Ecological regions are increasingly used as a spatial unit for planning and environmental management. It is important to define these regions in a scientifically defensible way to justify any decisions made on the basis that they are representative of broad environmental assets. The paper describes a methodology and tool to identify cohesive bioregions. The methodology applies an elicitation process to obtain geographical descriptions for bioregions, each of these is transformed into a Normal density estimate on environmental variables within that region. This prior information is balanced with data classification of environmental datasets using a Bayesian statistical modelling approach to objectively map ecological regions. The method is called model-based clustering as it fits a Normal mixture model to the clusters associated with regions, and it addresses issues of uncertainty in environmental datasets due to overlapping clusters.

1. INTRODUCTION

Ecoregions define recognizable areas which embody broad environmental and landscape structures. Ecoregion classification and subsequent boundary definition have a significant impact on natural resource management. The need for bioregionalisations was initially driven by conservation planning, but they have taken on extended roles as spatial units for tabulating environmental information (as opposed to socio-economic administrative units) and for the allocation of funding for the environment. In Australia a bioregional planning framework, called the Interim Biogeographic Regionalisation of Australia (IBRA) has been established [EA, 2000]. The biogeographical regions in IBRA are land areas comprised of interacting ecosystems that are repeated in similar form across the landscape. Typically the IBRA regions are based upon factors such as climate, lithology, geology, landforms and vegetation as surrogate indicators of the ecological processes that occur on land, particularly as relevant to conservation strategies and natural resource capability. The ecoregions are mapped at different scales within a hierarchy ranging from broad land types to local regional ecosystems (See Figure 1).

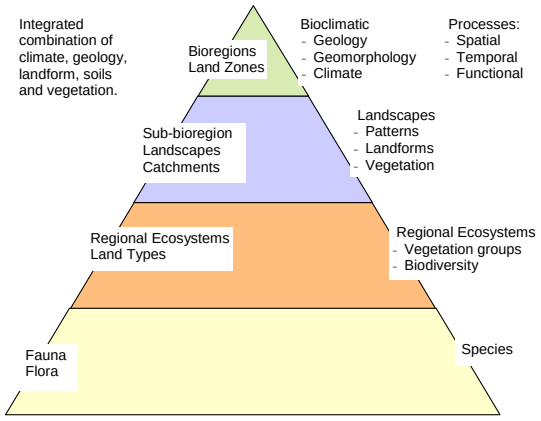

Figure 1. Conceptual hierarchy of bioregional classification at four levels. Adopted from Sattler and Williams [1999]

The focus of this paper is on sub-bioregions as areas of land that have a distinctive pattern of landform and vegetation which indicates major differences in land processes and biological communities [Sattler and Williams, 1999]. Sub-bioregions are mapped at a scale of 1:100,000. In Queensland, the delineation of sub-bioregions is largely overseen by an expert scientific panel who interpret available mapped information sources using their knowledge of the region. The regions are mapped as areas that have distinctive landscape patterns with permeable boundaries. With growing use of these regions within natural resource decision-making there is pressure to shift from subjective expert-based methods for defining bioregions to a more repeatable, scientifically defensible and objective system of classification. In response to this need a project was undertaken to make the expert input more explicit and to incorporate classification based on statistical analysis of geographic information. The guiding principle in the classification is to determine the key drivers amongst a range of abiotic environmental factors using cluster analysis to identify cohesive and separable classes from geophysical datasets.

The outline for the paper is as follows. The next section explains the location for the study area. Section 3 describes the Bayesian approach to classification. Section 4 describes the spatial and graphical tool used to elicit knowledge from experts that is used as prior information to guide the classifier. Section 5 illustrates the results for a classification. Section 6 summarises and discusses the significance of the work.

2. STUDY AREA

The results of the research are to be applied to eastern bioregions within the state of Queensland in Australia, however the paper will focus on one bioregion in the south-eastern corner of the state (Figure 2).

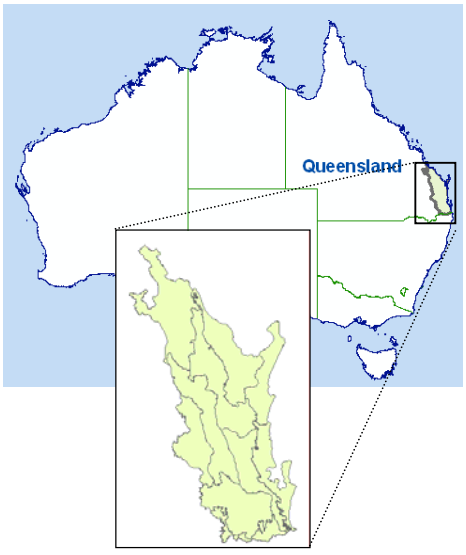

Figure 2. Locality map showing bioregion and sub-bioregions for the study region in South-East Queensland.

The bioregion covers 66,000 km^2 and comprises coastal plains, a major drainage basin for the

Brisbane river catchment, and mountain ranges. The area is sub-tropical and is considered one of the most species-rich and diverse parts of Australia for flora and fauna [Sattler and Williams, 1999]. There is significant settlement of the region with a population of approx. 2 million people, and the expectation this population will double in the next 40 years. Despite a number of national parks and smaller reserves the area has several vulnerable species that are endangered and bioregional planning plays an important part in decision-making for future development.

3. METHODOLOGY

Previous approaches to bioregionalisation have tended to be either expert-driven or data-driven [eg Bunce et al 2002, Hargrove and Hoffman 1999). For example the most recent set of Queensland's sub-bioregions [Sattler and Williams 1999] is based on expert opinion on sub-bioregional boundaries [see Morgan and Terry 1990]. As is common in these situations a Delphic approach was used, where a panel of several experts were consulted together about the location of boundaries, based on mapped and well-defined topographic features such as regional ecosystem boundaries (derived from aerial photography), ridgelines, etc [Neldner 2002]. In their assessments, experts also referred to other spatial information such as soils and climate. In contrast the most recent sub-bioregions for Tasmania [Peters and Thackway 1998] take a data-driven approach and make use of spatially extensive fine scale information both biotic and abiotic. This data was input to multivariate clustering techniques [Everitt and Hand, 1981], and then use this as input, post-hoc, to an expert panel process to address inconsistencies and other model inadequacies.

Here we propose a regionalisation approach that aims to balance inputs from both experts and data, integrated within a Bayesian statistical modelling framework. The basic premise [Congdon 2001] is that updating prior information on parameters using information provided by data (likelihood) provides posterior information on these parameters:

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood} \qquad (1)$$

This provides a natural framework for continually updating old models (priors) with new data (likelihood) as it arises to produce new improved models (posteriors). Readers are referred to Gelman et al [2004] for further information on Bayesian statistical modelling.

3. 1 Models

In many situations statistical distributions (eg Normal, Poisson, exponential, etc) do not fit the observed data. This is particularly true of environmental data where a mixture of environmental conditions could lead to different patterns in the data. Mixture models address this issue by explicitly allowing for a mixture of components, each described by a separate distribution, to combine together into an overall mixture distribution.

More precisely, we define a mixture distribution for K clusters or mixture components indexed $k = 1\ldots K$. Let w_k denote the weight or proportion of observations in each cluster. Denote by x the dataset with one row per observation and one column per variable (eg environmental attribute). Let θ represent the set of mixture model parameters. Then the overall mixture likelihood $p(\cdot)$ is defined as the weighted sum of mixture components $f(\cdot)$:

$$p(x \mid \theta) = \sum_{k=1}^{K} w_k \, f_k(x \mid \theta_k) \qquad (2)$$

A common choice for the model for each cluster is a multivariate Normal, giving rise to a Gaussian mixture model. In the k^{th} cluster, for the i^{th} observation on all variables x_i:

$$f(x_i \mid \theta_k) \equiv MVN_d(\mu_k, \Sigma_k) \qquad (3)$$

This indicates that each observation in the cluster is drawn from a multivariate Normal distribution of d dimensions with $d \times 1$ vector of means μ and $d \times d$ variance-covariance matrix Σ. This could mean for bioregionalisation that a particular region is defined by a 3D Normal distribution with mean rainfall 50mm pa, soil moisture 0.10, and elevation 50m. The standard errors could be, say, respectively 10mm pa, 0.04, and 20m, with the only non-negligible covariance being 42% between soil moisture and rainfall. If the variability of a variable is narrow (small standard error) then the cluster/region is closely linked to that environmental attribute. Similarly wide variance leads to little relationship between that geographical region and the environmental attribute in question. See Figure 3.

A method proposed by Dempster et al [1977] relies on introduction of extra (auxiliary) variables to facilitate the computations. These keep track of cluster membership for each observation. Let $z_i = k$ if the i^{th} observation falls into the k^{th} cluster. See figure 4. Then the weight of each cluster is just the same as the probability of cluster membership:

$$w_k = p(z_i = k) \quad (4)$$

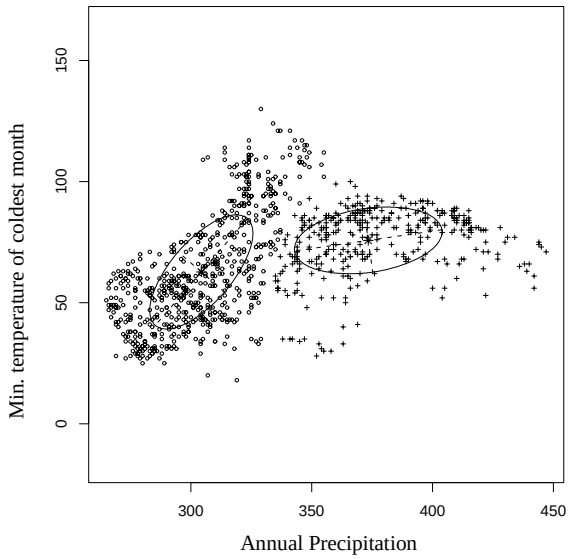

Figure 3. A mixture model with two variables and two clusters. The ellipsoids mark the standard deviation of the bivariate normal distribution that defines each cluster. The size, shape and orientation of a cluster's ellipsoid indicates the means, variances and correlations of the two variables for sites in the cluster.

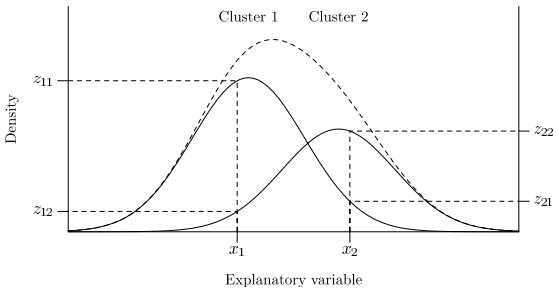

Figure 4. A mixture model in which the distribution of one variable is modelled as a mixture of the distributions of the variable at sites in each of two clusters. Site 1 is assigned to cluster 1 because the probability density at x_1 is greater for cluster 1 than that for cluster 2.

The computation of the mixture model is applied iteratively to explore the posterior distribution of each parameter repeatedly until most important parts of the distribution have been explored. This is the general idea behind Markov Chain Monte Carlo (MCMC) [Gelman et al , 2004]. Through MCMC we obtain dependent simulations that, once they've reached equilibrium, model the target posterior distributions (1) for each parameter. The challenge is to design an MCMC sampler that converges to equilibrium efficiently.

A general approach for implementing either Bayesian approach comprises three steps:

1. Designing priors
2. Designing MCMC samplers
3. Implementing MCMC

Designing the MCMC samplers and implementing the MCMC are not the focus of this paper. We focus on the first stage of designing appropriate priors in the next section.

3. 2 Priors and eliciting expert knowledge

In equation (1) the priors and likelihood have equal impact. The theoretical mixture model likelihood is defined through equations (2)-(4). Designing appropriate priors is somewhat of an "art" and requires two main stages:

1. Select appropriate priors to enable dialogue with expert(s) so that they can describe their prior knowledge in a form suitable for input to the model.
2. Design and implement elicitation processes (experiments) to quantify the prior knowledge held by experts.

These two stages are closely linked. Without knowing the form of the prior the elicitation process at worst can provide irrelevant information. On the other hand, without a rigorous elicitation experiment, it is difficult to ensure the validity, repeatability and transparency of priors obtained.

For Gaussian mixture models there are four main types of priors we can consider, depending on the type of expert knowledge available.

Expert knowledge	Appropriate prior
Experts know nothing (objective ~ Frequentist)	Non-informative (improper) priors
Experts know something about model coefficients (means and variances on each variable in each cluster)	Informative conjugate
	Informative semi-conjugate
Experts know something else	Data Augment-ation priors

We focus on the second more usual choice, the informative conjugate prior: informative since prior knowledge on means and variances in each cluster informs the model (has impact on results), and conjugate since the choice of prior distribution factors out "nicely" mathematically. For the Gaussian mixture model, this prior comprises a Normal distribution for cluster means conditional on known cluster variance ($\mu_k \mid \Sigma_k$ in Equation 5), with an inverse Wishart distribution for the inverse

covariance matrix (Σ_k^{-1} in Equation 5) [Diebolt and Robert 1994].

$$\mu_k \mid \Sigma_k \sim N(m_k, s_k) \quad \Sigma_k^{-1} \sim W^{-1}(\nu_k, \varphi_k) \quad (5)$$

Each prior has a number of hyperparameters m_k, s_k, μ_k, ν_k describing respectively the best guess of the value and precision of the cluster means, and best guesses for cluster covariance matrix and the "effective" amount of prior information used to derive these.

These priors match expert knowledge about average and standard deviation of each environmental attribute within each cluster, where the mean depends on the standard deviation.

4 SPATIAL ELICITATION TOOL

Eliciting information from experts for input into Bayesian models requires a blend of psychological survey design skills, designing questions for interview, determining who is interviewed and how many times. The challenge is that instead of factual information, we require knowledge as synthesized by the expert to be deconstructed and quantified in a form like (5) suitable for input into modelling. These issues are addressed in the expert elicitation literature [O'Hagan 1988].

4.1 Design

To this end we have designed a computer assisted elicitation process that uses a spatial and graphical tool to help the user visualize and explore the data. Essentially the user can interact with data from various "viewpoints" each with a different activity:

- *Cartographic*: select an existing sub-bioregion or select attributes to spatially define a "new" sub-bioregion,
- *Data exploration*: inspect and adjust histograms of each environmental attribute,
- *Spatial analysis*: map several environmental attributes within the geographic region.

Thus a user can choose to define a sub-bioregion: in geographic space as a cartographic view or in variable space as an environmental "domain". The aim is to elicit the priors (μ_k , Σ_k) for each cluster, where a cluster corresponds to a sub-bioregion. The two step process in eliciting these priors is explained below.

Use the cartographic view to geographically select areas that characterise each sub-bioregion. This is typically specified in terms of land classes for vegetation types, landforms and species distributions. For example, an ecologist may select areas that form a bio-region made from coastal lowlands with Banksia open forest. This is carried out in a GIS with custom tools to assist in making attribute selections. The geographical selections are used to analyse environmental datasets and extract variables within the selected regions.

The data exploration view shows histograms for the set of environmental variables within the above geographical selection. These variables are the key abiotic factors used to classify and differentiate between sub-bioregions. For example, continuous variables for climate, topography and soil characteristics. An important facet of the data exploration view is the ability to define thresholds for variables. For instance the user may clearly want to eliminate a certain range of values for a variable (eg particular soil qualities or low rainfall values). These thresholds can be defined in a univariate or bivariate fashion. A graphical tool invoked from the GIS shows adjustable histograms of several environmental attributes within that region. This provides hyper-prior parameter estimates m_k,s_k for the mean. Instead of eliciting the covariance matrix from the user, we use a sample covariance matrix φ_k estimated from a sub-sample for that region. The user can adjust another control to set the degrees of freedom (or effective prior sample size) ν_k to reflect certainty in this matrix. The graphical tool has functions to store the prior estimates of mean (best guess and certainty) as well as the degrees of freedom of the covariance matrix for each sub-bioregion to be classified. This "experimental" data along with other basic metadata is used as priors in the Bayesian cluster classification.

4.2 Implementation and Visualisation Interface

A map-based user interface has been developed using GIS technology to display parameters as maps and charts. See Figure 5. In the data exploration view, means are visualized by splitting the x-axis on histogram and slider bar into three colours. Data symbolization is based upon a variation of a boxplot to show where credible intervals are which are then displayed on the map view [Car et al, 1999]. The slider control allows a user to adjust class breaks interactively, and these changes are automatically reflected in the colours displayed on the chart and the map cartographic view. This provides an effective means for a user to interactively set an estimated credible interval for single variables within the region.

This information gathered from experts is feed into the informative priors and is recorded as part of an experimental workbook. Elicitation information includes the name of the expert, date, remarks, centre value and bounds for each variable analysed. This information may be used to weight (e.g. based upon certainty or expert knowledge)

informative priors and to document the results of a classification.

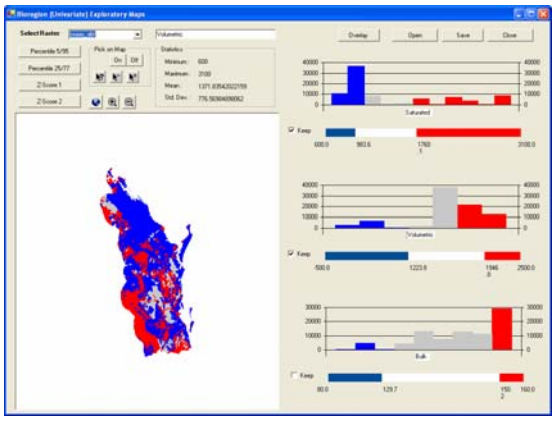

Figure 5. Univariate data visualization.

The graphical interfaces includes a map-based (cartographic) view and a graph-based (data exploration) view. The user may add up to three environmental variables as maps. These two views are linked so that changes in one are simultaneously reflected in the other. Up to three maps may be added in this way, and then a combined map or overlay may be created to see the mapped overlap distributions. The overlap distribution is representative of the confidence intervals of the means of the environmental attributes of a sub-bioregion. The expert can then interact with the map to add or remove areas from the sub-bioregion. Hence, the cartographic view and exploration view are dynamically linked. The expert may view another graphical interface for exploration of combinations of the variables. A map and a scatter diagram with a background density frequency chart are displayed for combinations of the two intervals attributes in two dimensions. The co-occurrence of related variables show up as clusters which the expert can refine by selecting the representative center or mean μ in two dimension, and an area around this center representing the credible interval. This information is also saved as prior information for the classifier.

5. RESULTS

The approach may be validated visually against the existing sub-bioregions by fitting mixture models to the most significant environmental variables and a comparison made to see what adjustments are suggested by the resulting clusters. Figure 6 shows the results of this computation for south-east Queensland and it is seen that the adjustments are minor. The most significant variables used in the analysis were selected statistically with a dimension reduction technique. In our south-east Queensland case study we were able to adequately fit mixture models to the existing sub-bioregions with a manageable number of topographic, climate and soil variables. Conformance with the existing sub-bioregions could be effectively controlled by manipulation of the relative weights placed on the priors and data variables (Figure 6).

(a) (b) (c)

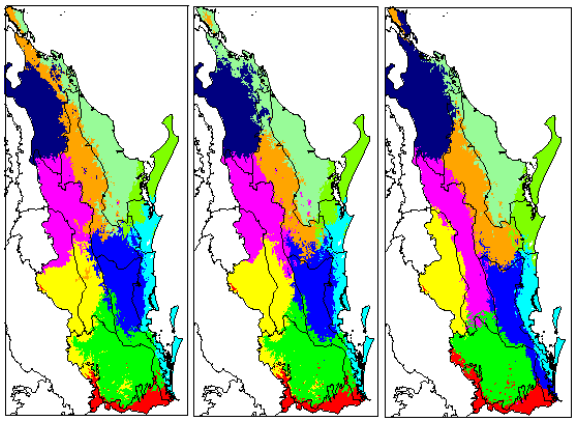

Figure 6. The existing sub-bioregions for south-east Queensland shown by solid lines on Bayesian mixture model classifications with: (a) no prior, (b) moderate weighting on priors, and (c) strong weighting on priors. The priors were calculated from the existing sub-bioregions.

6. CONCLUSION

The significance of the research is that a Bayesian approach allows us to combine qualitative information and quantitative data in classification. Hence combining - the previously competing - approaches of expert panel and data classification. Bayesian mixture models provide a method for classifying ecoregions with a formal statistical procedure that fits overlapping clusters. When mapped spatially the cluster components relate well to coherent bio-regions. This is illustrated in Figure 6, which also shows the results of adjusting the relative weightings on expert knowledge and data between bio-regions. Our elicitation tool enables experts to interactively specify quantitative model parameters (e.g. means and covariance matrices) by viewing and manipulating familiar entities such as maps and histograms. The results are presented visually in this paper; future work will provide details on model diagnostics, model performance, and model comparisons.

7. REFERENCES

Environment Australia, Revision of the Interim Biogeographic Regionalisation for Australia (IBRA) and Development of Version 5.1.

Summary Report, Canberra, Environment Australia, 2000.

Bunce, R., P. Carey, R. Elena-Rossello, J. Orr, J. Watkins and R. Fuller, A comparison of different biogeographical classifications of Europe, Great Britain and Spain, *Journal of Environmental Management* 65, 121-134, 2002

Carr, D., A. Olsen, S. Pierson and J-Y. Courbois, Boxplot Variations in a Spatial Context. Statistical *Computing & Statistical Graphics Newsletter*, American Statistical Association, 1999.

Congdon, P. *Bayesian Statistical Modelling*, Wiley, New York, 2001.

Dempster, AP., N. Laird and D. Rubin, Maximum likelihood from incomplete data via the EM algorithm (with discussion), *J Roy Statist Soc Ser B*, 39: 1-38, 1977.

Diebolt, J., and C. Robert, Estimation of finite mixture distributions through Bayesian sampling, *J. Roy. Statist. Soc. Ser. B*, 56(2), 363-375, 1994.

Everitt, B., and D. Hand, *Finite mixture distributions*, London: Chapman and Hall, 1981.

Gelman, A., J. Carlin, H. Stern, and D. Rubin, *Bayesian Data Analysis*, 2nd edition. Chapman and Hall/CRC, Florida., 2004

Hargrove, W., and F. Hoffman, Using multivariate clustering to characterize ecoregion borders, *Computing in Science and Engineering* 1(4), 18-25, 1999.

Morgan, G., and J. Terrey, Natural regions of western New South Wales and their use for environmentl management, Proc Ecol Soc Aust 16, 467-473, 1990.

Neldner, VJ. *Summary of procedure for creating regional ecosystem maps as defined under the Vegetation Management Act 1999*. Technical Report. Brisbane: Environmental Protection Agency, 2002.

O'Hagan, A., Elicitation of expert beliefs in substantial practical applications. *The Statistician* 47(1), 21–35, 1998.

Peters, D., and R. Thackway, *A New Biogeographic Regionalisation for Tasmania*. Hobart, Tasmanian Parks and Wildlife, 1998

Sattler, P. and R. Williams, *The Conservation Status of Queensland's Bioregional Ecosystems*, Environmental protection Agency, Queensland Government, 1999.

8. ACKNOWLEDGEMENTS

We thank Petra Kuhnert and Robert Denham for helpful early discussions, ideas and suggestions. This work was supported by ARC-SPIRT Grant C00107484 and by industry partner Queensland Environmental Protection Agency, Australia.