Bioregion Classification Using Model-Based Clustering: A Case Study In North Eastern Queensland

A. Accad[a], S. Low Choy[b], D. Pullar[c] and W. Rochester[d]

[a] The Queensland Herbarium, Environmental Protection Agency, [b] Bayesian Research & Applications Group School of Mathematical Sciences Queensland University of Technology [c] Geography Planning and Architecture, The University of Queensland, [d] CSIRO Marine Research , E-Mail: arnon.accad@epa.qld.gov.au

Keywords: *biogeography; bioregions; subregion; statistical modelling; GIS; finite mixture models; clustering*

EXTENDED ABSTRACT

The Interim Biogeographic Regionalisation of Australia (IBRA; Environment Australia, 2000) is a planning framework defining land areas comprised of interacting ecosystems repeated across the landscape. In many states these bioregions are currently arrived at by consensus of an expert panel (Neldner *et al*. 2004) and well accepted as a spatial unit for planning and environmental management. This work was motivated by the need 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 present bioregional boundaries of Queensland version 4.3 are shown in Figure 1. The case study is situated in North Eastern Queensland in an area where a boundary change was proposed. This research investigates at the meso-scale the success of broad climate and soil variables in identifying patterns and processes. We report results based on three bioclimate and soil variables, suggested through exploratory analysis and model sensitivity: temperature seasonality (bc04); annual precipitation (bc12); and B horizon available water holding capacity (baw).

This paper compares a range of statistical methods for bioregion classification, within a continuum of data-driven to expert-driven, including Bayesian methods. Model-based clustering moves away from traditional methods which delineate boundaries, instead assessing similarity between and within geographic regions and environmental envelopes. Bayesian statistical modelling enables explicit input of expert prior knowledge during development of bioregions. We assessed two alternative prior knowledge bases: vegetation communities or existing bioregion boundaries. In data poor areas expert defined boundaries are feasible, but subjective. Vegetation-based priors can be considered more objective, although they require subjective identification of communities, and are a useful alternative to expert boundaries.

This study confirmed that experts contribute knowledge beyond what is currently mapped on bioclimate and soils. The Bayesian model-based approach has significant benefits in assessing impact of different types of expert knowledge for bioregions—either mapped communities or boundaries—as well as for quantifying precision of modelled regions.

Practically we found that the Frequentist model-based approach was useful in initial stages of modelling. The distance-metric based approaches to clustering though relatively simple to implement provide qualitatively different boundaries, and require an unwieldy process for obtaining predictions, for which no assessment of uncertainty is available.

For bioregionalisation of new areas, expert-defined boundaries may still play a role, although this has now been demonstrated to be more useful when combined with bioclimate and soils datasets in a Bayesian framework. In data-rich areas, the Frequentist model-based approach may suffice.

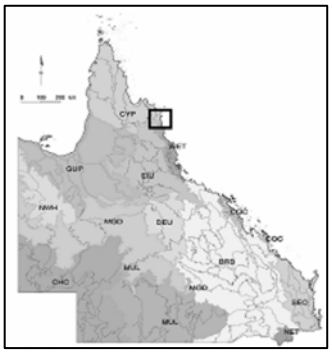

Figure 1. Queensland bioregions and study area.

1. INTRODUCTION

The bioregions of Queensland provide a framework for conservation assessment and protection. They support a systematic approach to conservation, essential if biodiversity is to be effectively protected (Sattler and Williams, 1999). To this end a two level biodiversity hierarchy of protection is required, at landscape scale via bioregional and subregional protection strategies, and finer scale ecosystems, species and genotype protection via Regional Ecosystems and land types protection. Bioregional and subregional mapping also provides a natural reporting and decision-making framework for land and vegetation management statewide. It supports application of the *Vegetation Management Act (VMA), 1999*, which assigns vegetation communities and regional ecosystems (REs) a vegetation management status according to the percent of the pre-clearing extent which remains within bioregions.

The subregions considered in this study are at the centre of a proposed boundary change and include the Daintree-Bloomfield, Hodgkinson Basin and Starcke Coastal Lowlands. These subregions form the bioregion boundaries between the Wet Tropics (WET), Einasleigh Uplands (EIU) and Cape York Peninsula (CYP) bioregions of North Queensland (Fig. 2).

This paper applies a new model-based approach (Rochester *et al.* 2004; Low Choy *et al*. 2005; Pullar *et al*. 2005) including both data and expert knowledge in a Bayesian statistical model to describe and assess regional boundaries. By focusing on this current issue of redefining a bioregional boundary in the case study, we may compare a range of statistical methods for bioregion and subregion classification, within a continuum of data-driven to expert-driven methods.

2. METHODS

2.1 Ecological framework and the Expert approach

The ecological model proposes that subregions are shaped by their broad landforms, climate and vegetation. This hypothesis is to be tested in this study. The ecological model (Table 1) refined in consultation with experts (Neldner *et al*. 2004) illustrates the complexity of the study area, and highlights that the current bioregional boundaries are positioned in a location of steep environmental gradients, which reflect reduction in elevation and rainfall, and a shift from close rainforest through

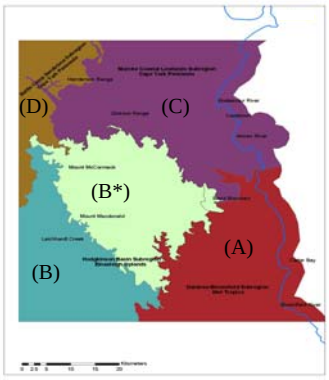

Figure 2. A. Daintree-Bloomfield, WET; B. Hodgkinson Basin, EIU; C. Starcke Coastal Lowlands and D. Battle Camp Sandstone, CYP.

woodlands to low woodlands. The subregions in the area reflect a large amount of within subregion variability regarding landforms and as a result, diverse types of vegetation species and communities.

Botanists advised that the current Einasleigh-Cape York boundary should be reviewed. Their advice was that the proposed boundary should be amended to reflect the distribution of vegetation communities containing key species *Corymbia nesophila* on metamorphics and *Eucalyptus tardecidens* (map units 82 and 142 respectively Neldner and Clarkson, 1995; Fig. 3; Pers. Comm. Addiccott, 2005). This information assisted the experts in splitting the Hodgkinson Basin subregion EIU into two areas B and B* (Fig. 2). This prompted a proposal to the Bioregions expert panel to assign the northeastern area (B*) to the Starcke Coastal Lowlands subregion CYP.

Table 1. The ecological conceptual model.

Name	Elevation	Rainfall	Vegetation
Daintree-Bloomfield of the Wet Tropics (**WET**)	> 1000m	> 3000mm in Carbine and Thornton-1600mm in Windsor	Rainforests , vine thickets and Sclerophyll woodland and forest .
Battle Camp Sandstone of the Cape York Peninsula (**CYP**)	Mean elevation above 200m with max elevation below 600m	> 1600mm rapidly decreasing to less than 1000mm in the west	Woodlands
Hodgkinson Basin of the Einasleigh Uplands (**EIU**)	Upper catchment of the Mitchell and Gilbert Rivers	Rain shadow < 1500mm	Low woodlands on loamy lithosols
Starcke Coastal Lowlands of the Cape York Peninsula (**CYP**)	Lowland areas below 100m with peaks reaching 500m	> 1400mm	Woodlands, low open woodlands and Heathlands

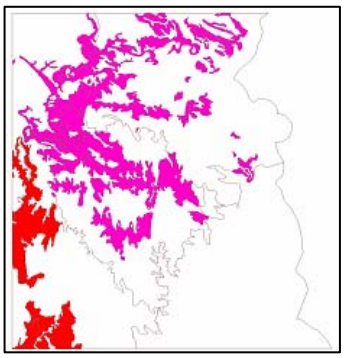

Figure 3. The ecologists' use of vegetation community distributions as a means to propose a new bioregion boundary**.**

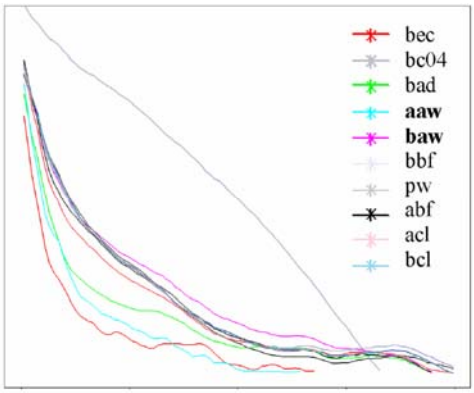

Figure 4. Correlogram comparing spatial auto-correlation in some competing and correlated soils attributes identified after dimension reduction. (For variables assessed see Rochester *et al*. 2004).

2.2 The Data Model

The data model (Austin, 2002) consists of the statistical design decisions made regarding how the data are collected and measured. The data model in the study area operates on 1ha square grid cells. This scale was considered appropriate for reflecting subregional differences in this study area and did not overstate the spatial accuracy of the datasets. The design for subsampling from the spatially extensive datasets was vital for ensuring representativeness and eliminating pseudo-replication arising from sampling sites too close together. A stratified random sampling process was applied to spatially subsample variables at a set of 2000 locations, stratified by vegetation types to ensure that environmental variation relevant to subregions was represented. Correlograms (Fig. 4) confirmed that this sample size avoided pseudo-replication.

Based on the ecological model, the data model addresses three themes considered to be the drivers of environmental variation in the study area, namely geomorphological, bioclimatic and soil properties. Recently available information provided over 100 spatial datasets that could be indicators of differences between the four subregions. Initial selection of variables was achieved using exploratory data analysis techniques including univariate and multivariate assessments: maps to view spatial pattern, histograms to suggest transformations; MANOVA to summarize within and between subregion variation of variables; k-means clustering, Factor analysis and correlation matrices annotated with hierarchical agglomerative clustering helped select variables from correlated groups. After transformation all variables were scaled by subtracting the mean and dividing by the standard deviation to help in comparisons between variables in later diagnostics. The ecological model was used at each stage to ensure that variables selected were ecologically justified.

Statistical distributions of variables were compared across subregions (Fig. 5) via boxplots. This highlighted variables which differ greatly in distribution (eg bc04 temperature seasonality, bc12 annual precipitation) across subregions or do not discriminate between subregions (eg aspect). Whilst some variables may explain differences between subregions, their selection should also depend on the required scale. For example the soil B horizon available water holding capacity (baw) was selected over the A horizon (aaw) counterpart as baw was changing slower than aaw over spatial distances and as a result will explain differences between rather than within subregions, and thus provide less fragmented clustering (Fig. 4).

2.3 Statistical techniques

Statistical modelling involves specification of the statistical technique, including assumed distributions, together with the decision-making framework, such as hypothesis tests (Austin, 2002; Gelman *et al*. 2005). We consider two main types of statistical techniques: heuristics, namely distance-metric based multivariate techniques; as well as model-based approaches based on finite mixtures of distributions (*eg* Gelman *et al*. 2005). Heuristic approaches (Hartigan, 1975) focus on allocating sites to regions which minimize a cost function for dissimilarity between regions. Modelling approaches describe a multivariate distribution of environmental attributes that are useful predictors for the unknown allocation of sites to subregions, in this case the four subregions on the boundaries between the WET, EIU and CYP bioregions. Both Frequentist and Bayesian implementations of the finite mixture model are investigated. Traditional bioregional boundary mapping uses similar data—geological maps, climate information, image interpretation and site data information (Neldner *et al*. 2004)—within a qualitative approach addressing the same

ecological model, but integrated conceptually by experts in that they describe boundaries which indirectly reflect the distribution of attributes in each region.

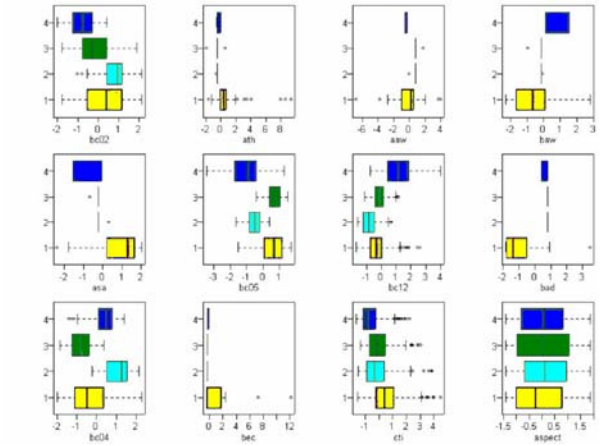

Figure 5. Statistical distribution of attributes in subregions WET (Blue), EIU (Yellow), WCYP (Aqua), ECYP (Green).

Here the heuristic approaches represent the purely data-driven or data mining approaches (Hastie *et al*. 2001). Frequentist model-based clustering using finite mixtures provide a data-driven approach based on expert-defined hypotheses, and can be extended to the Bayesian context via prior models for parameters. These prior models provide a mechanism for explicitly balancing input from both experts and data, from the outset. The Expert-driven approach based on Delphic consultation of panels of experts (Neldner *et al*. 2004) refers to data but relies on more subjective integration of this information.

Distance-Metric Based Cluster Analysis

Cluster analysis was the simplest method considered to group sites and, for a specific distance metric, measures dissimilarity between sites (Hartigan, 1975). For example agglomerative clustering continues to aggregate groups together according to increasing dissimilarity until there is just one group. The number of groups was selected *posthoc* (Hastie *et al*. 2001). Heuristic approaches require imputation of a simplistic model in order to extrapolate predicted site allocation across the whole region. We found classification trees using recursive partitioning (CART in RPart package in S-PLUS, Venables and Ripley, 1994) to work well although other methods such as linear or quadratic discriminant analysis could also have been used. Limitations are the arbitrary algorithm settings (distance metric, stopping criterion), and no model basis for prediction and hypothesis testing.

Model-based clustering

First consider the unrealistic problem where we already know allocation of sites to subregions. Then environmental variables X in each subregion k can be modelled as following a multivariate Normal (MVN) distribution.

$$\phi(X_i \mid \mu_k, \Sigma_k) = (2\pi)^{-p/2} \mid \Sigma_k \mid^{-1/2} \exp\left\{-\tfrac{1}{2}(X_i - \mu_k)\Sigma_k^{-1}(X_i - \mu_k)\right\} \quad (1)$$

This can be addressed through a straightforward discriminant analysis. For bioregionalisation however site allocation is unknown. Define $z_i = k$ if the ith site is allocated to kth subregion and let $Pr(z_i = k) = w_k$, where w_k is the overall weights of the kth subregion. Thus the overall likelihood model is a finite mixture of MVN distributions:

$$p(X,z|\mu,\Sigma)=\prod_{i=1}^{n} w_k p(X_i|z_i=k,\mu_k,\Sigma_k) \quad (2)$$

Of key interest, esp. for bioregionalisation is prediction of site allocations z. Clusters or environmental envelopes (sites with similar variables X) can be mapped to subregions in geographical space (sites located close together), with clusters/subregions having environmental variables centred at means μ_k. Covariance matrices Σ_k have eigenvalue decomposition

$$\Sigma_k = \lambda_k D_k A_k D_k^T \quad (3)$$

where for each cluster k, D_k is the orientation of the environmental envelope with respect to the variable axes, defining principal components (PCs), A_k is proportional to the eigenvalues of Σ_k and determines the shape of the environmental envelope which reflects impact of each PC and therefore of variables, while λ_k is a scalar that determines the envelope's volume and reflects the subregion's overall environmental variation. Allowing the orientation (PCs), shape (variable impact) and volume (heterogeneity) to be the identity, equal or vary across clusters (subregions) leads to variance models of differing complexity (Bensmail *et al*. 1997; Low Choy *et al.* 2005).

Frequentist approach

The Mclust package for S-PLUS (Fraley and Raftery, 2002) implements Frequentist estimation for this finite mixture of multivariate Normal (FM MVN) distributions, via the EM, (Expectation-Maximization) algorithm. The calculation of uncertainty associated with site allocation z for this model is easily estimated from model coefficients (Low Choy *et al*. 2005). Resampling is required to assess precision of estimated means and covariances. Best performance, ie number of

subregions or variance parameters, corresponds to maximum BIC (Fraley & Raftery, 2002).

Bayesian approach

With the Bayesian approach we derive a prior model for all parameters, $\{z_i, \mu_k, \Sigma_k, w_k\}$, is then combined with the data model (1)-(3), to obtain updated (posterior) estimates of parameters.

$$\text{Posterior} \propto \text{Prior} \times \text{Likelihood} \qquad (4)$$

Here priors are based on initial estimates of subregional boundaries. We use conjugate dependent priors: Dirichlet for cluster weights w, inverse Wishart for covariances Σ_k and MVN for the mean given the variance $\mu_k|\Sigma_k$. For expert-defined subregions (developed from viewing different data), the mean and covariance matrices for modelled variables can be estimated from sample estimates within each subregion. Prior knowledge was strong in the highly informative situation where expert knowledge is assigned high precision, to weak for the nearly non-informative situation where it has low precision (for details see Low Choy *et al*. 2005).

Two different sources of prior knowledge are investigated. One is the existing set of subregional boundaries, an integrated set of expert-defined boundaries where various environmental attributes may have been used to define different boundaries. The other is the spatial distribution of vegetation communities in different areas that was thought to be indicative of the separation between lower and upper Hodgkinson Basin EIU (Fig. 3). These two sources are quite different. Vegetation communities result from expert selection of communities from knowledge applied at a finer scale, via delineation of each vegetation polygon. Expert-defined subregion boundaries require experts to integrate knowledge at a broader scale across several attributes.

Model Evaluation

Explanatory ability could only be assessed for a model-based approach, where we seek accuracy in estimating parameters namely subregional extent w_k, means μ_k and covariances Σ_k. This could only be achieved posthoc using resampling techniques for the Frequentist approach so instead we use a Bayesian approach to estimate direct uncertainty via credibility intervals. Internal model fit to the data was assessed for models via Schwarz's BIC to check trade-off between covariance parameterization and number of subregions. Note that weaker prior based on expert knowledge will automatically show closer fit to data using BIC. Predictive ability in site allocation was assessed via visual inspection of mapped predictions or equivalent concordance statistic, and by using MANOVA. Predictive uncertainty can be estimated by the probability that the site was allocated to any other subregion given the model. See Low Choy *et al.* (2005) for details.

3. RESULTS

3.1. Site allocation maps

Cluster Analysis

The results of applying hierarchical clustering, with Euclidean distance metric and Ward's linkage, extrapolated to region via classification trees, illustrate that a very narrow area of the West WET Bioregion resembles EIU and CYP (Fig. 6). Similarly the two CYP subregions are found to be quite similar with an area in Starcke Coastal Lowlands CYP resembling the Hodgkinson Basin EIU. Some small areas in the results are shown as outliers of other subregions. These areas may have been too small or too far to form a subregional outlier when the experts considered the delineation of the subregions.

Frequentist Model-based clustering

Predicted site allocation (Fig. 7a) using the Frequentist model have uncertainty (Fig. 7b) which is highest on the boundary between CYP and EIU bioregions. This confirms the source of current debate regarding the boundary. High uncertainty also occurs for some disjunct outliers, also identified using agglomerative clustering, that share soil and bioclimate profile in common with geographically distant areas. Since small, they can be merged with their surrounds for most purposes. Similar to agglomerative clustering a narrow area of the Western WET resembles EIU and CYP. A narrow strip of the northern EIU resembles the Starcke Coastal Lowlands CYP.

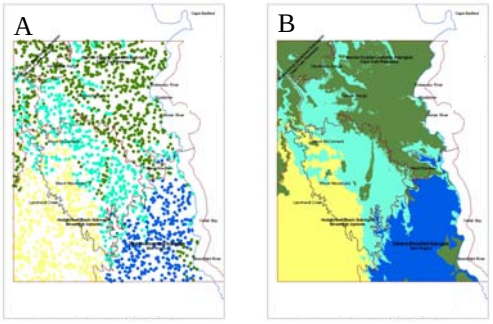

Figure 6. (A) Results from agglomerative clustering (BIC not applicable). (B) Extrapolation via CART.

Bayesian Model-based clustering

With moderate emphasis on a vegetation community prior, subregions site allocation are

most certain where the two key communities do not occur, and most uncertain in northern EIU (B* in Fig. 2). Similar to the Frequentist model, a boundary in WET is delineated which joins Battle Camp Sandstone CYP with northern Hodgkinson Basin EIU (Fig. 8). Not surprisingly the Bayesian model, with strong emphasis on current expert-derived boundaries, obtains the closest match to them (Fig. 9a). The largest difference is assigning Hodgkinson Basin to the Starcke Coastal Lowlands CYP, which is uncertain. All boundaries show high uncertainty (Fig. 9b).

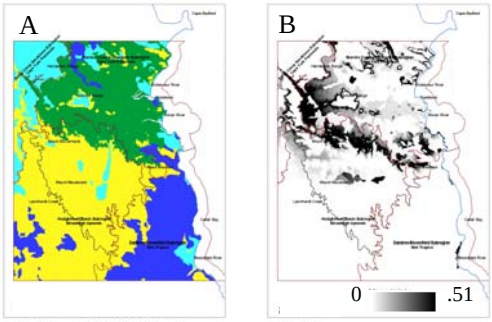

Figure 7. (A) Results from Frequentist model, (BIC = -12,764). (B) Site allocation uncertainty.

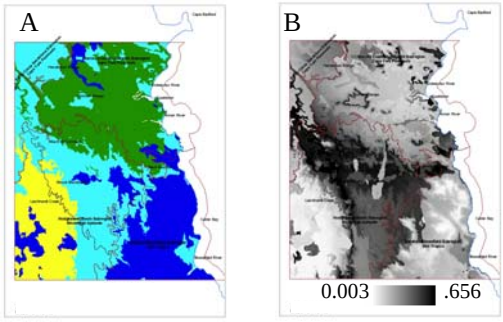

Figure 8. (A) Results from Bayesian model with a moderate vegetation prior, (BIC = –22,719). (B) Site allocation uncertainty.

3.2 Model parameter assessment

Comparing distribution of annual precipitation (bc12) for each set of modelled subregions (Fig. 10), we see that WET has a much tighter range via agglomerative clustering and a higher and wide range via Frequentist modelling. Bayesian modelling with existing subregion prior shows tighter modelled rainfall distribution in all regions apart from WET. WET has the highest range of rainfall for all models apart from the Bayesian model with vegetation prior. Models disagree which of the CYP subregions (east and west) have lowest rainfall: agglomerative clustering suggests west is lower; in contrast Bayesian with vegetation prior suggests east is lower, and other three models suggesting some overlap. Bayesian modelling with both priors shows precise estimates of mean rainfall (tiny "waistband") in all regions. Waistbands on boxplots (Venables & Ripley, 1994) for Bayesian results indicate precision of quantiles of fitted distributions.

3.3. Model fit to data

Not surprisingly data-driven Frequentist results show much closer fit to data, as BIC is almost twice that of Bayesian models employing expert knowledge as priors. However of the two Bayesian models, strong weight on an expert-defined subregion boundary prior results in a marginally closer fit to data.

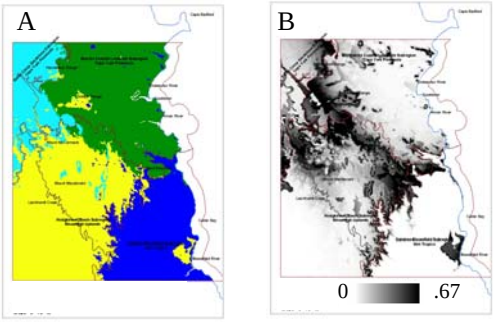

Figure 9. (A) Results from Bayesian model with strong prior based on existing subregions (BIC = –21,188). (B) Site allocation uncertainty.

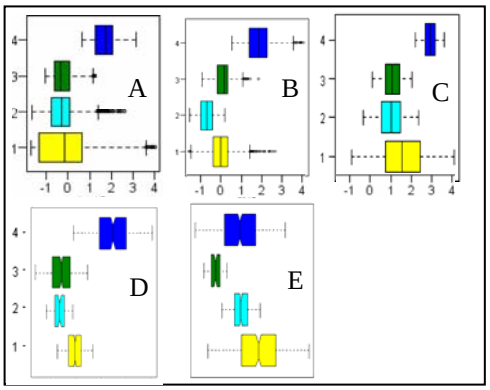

Figure 10. Annual precipitation (bc12) compared across models: A. Current expert boundary; B. Agglomerative; C. Frequentist; D. Bayesian, strong bioregion prior; E. Bayesian, moderate vegetation prior. Subregion are WET (Blue), EIU (Yellow), WCYP (Aqua), ECYP (Green).

4. DISCUSSION

The modelled distribution of annual precipitation (Fig. 10) demonstrates credible differences between subregions, supporting the initial ecological hypothesis. Since bioregions are an abstract concept there is no gold standard for assessing predictions. Nevertheless mapped site allocations (Figs. 6-9) show more support for

existing boundaries for Frequentist or Bayesian expert boundary prior models, but more support for the proposed boundary (Fig. 2) under the Bayesian vegetation prior. The heuristic suggests a boundary approx. halfway between. Also, when we vary the precision given to priors in Bayesian models gradually, we find that expert-defined boundaries only have a strong influence on results when upweighted ten times. Moreover expert knowledge in most cases tightens modelled distributions of bioclimate and soil.

5. CONCLUSIONS

This study confirmed that experts contribute knowledge beyond what is currently mapped for bioclimate and soils. Further study and mapping will be required before we may determine exactly what this extra knowledge represents. The Bayesian model-based approach has significant benefits in assessing impact of different sources of prior knowledge for bioregions as well as for quantifying precision. The Bayesian results from two different priors suggests that future research may integarte expert prior knowladge from different sources and scales, eg. integate the expert panel derived landscape subregional boundaries with the finer scale vegetation communities. Other extensions to be explored include variables selection and the number of subregions.

Practically we found that the Frequentist model based approach was useful in initial stages of modelling, with faster assessment of model sensitivity to variables and number of subregions. Heuristic approaches to clustering though relatively simple to implement provide qualitatively different boundaries, and require an unwieldy process for obtaining predictions, for which no assessment of uncertainty is available.

6. REFERENCES

Austin, M, 2002, Spatial prediction of species distribution: an interface between ecological theory and statistical modelling, Ecological Modelling 57(2-3), 101-118.

Bensmail, H., Celeux, G., Raftery, A. E. and Robert, C. P. (1997), 'Inference in model-based cluster analysis', Statistics and Computing 7, 1–10.

Ellison. A.M. 1996, An introduction to Bayesian inference for ecological research and environmental decision-making. *Ecological Applications* 6(4), 1036-1046.

Environment Australia, 2000. Revision of the interim biogeographic regionalisation for Australia (IBRA) and development of version 5.1. Summary Report, Environment Australia, Canberra.

Fraley, C. and A. E. Raftery (2002) "MCLUST: Software for model-based clustering, density estimation and discriminant analysis", Technical Report, Department of Statistics, University of Washington. http://www.stat.washington.edu/mclust

Hartigan, J.A. (1975). Clustering Algorithms. New York: John Wiley and Sons.

Hastie, T, Tibshirani, R. and Friedman, J. (2001) "The Elements of statistical learning: Data mining, Inference, and Prediction", Springer-Verlag: New York.

Kynn, M. 2005, Eliciting-Expert knowledge for Bayesian logistic regression in species habitat modelling. PhD thesis; QUT.

Low Choy, S. J, Rochester, W. (2005). A Bayesian model-based approach combining expert knowledge and GIS data for subregionalisation of Queensland's terrestrial bioregions. In preparation.

Neldner, V.J, Wilson, B.A, Thompson, E.J. and Dillewaard H.A. (2004) *Methodology for Survey And Mapping of Regional Ecosystems and Vegetation Communities in Queensland.* Version 3.0. Queensland Herbarium, Environmental Protection Agency, Brisbane.

Neldner V.J. and Clarkson J.R. (1995). 'Vegetation survey and mapping of Cape York Peninsula'. Cape York Peninsula Land Use Strategy, Office of the Coordinator General and Queensland Department of Environment of Heritage, Brisbane.

Pullar D., Low Choy, S. J., Rochester, W., Accad, A., Williams, K. and Neldner, J. (2005) Data analytic methods and expert knowledge in ecoregion classification submitted for publication.

Rochester, W, Accad, A, Low Choy, S, Neldner, V, Pullar, D, and Williams, K. 2004. Final Report UQ-EPA Subregion Classification Project. The University of Queensland, Brisbane, Australia.

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

Vegetation Management Act, 1999. Queensland Government Act No. 90 of (1999). http://www.legislation.qld.gov.au/LEGISLTN/CURRENT/V/VegetManA99.pdf .

Venables, W.M. and Ripley, B.D. (1994). Modern applied statistics with S-Plus. Springer-Verlag: New York.