ATLAS OF LIVING AUSTRALIA

REPORT ON THE

SPATIAL ANALYSIS TOOLKIT WORKSHOP

3RD & 4TH DECEMBER, 2009

Paul Flemons
Ben Raymond
Peter Brenton
Lee Belbin

Author(s):

Version: V2.0

Date: 2nd January 2010

Last updated 19/02/2010 2:23:00 PM

Revision History

Version	Date	Author(s)	Change description
0.1	7/12/09	P. Brenton	Initial draft – Insert & re-draft personal workshop notes
0.2	9/12/09	P. Flemons	First working draft incorporating overview
0.3	11/12/09	P. Flemons, B. Raymond, L. Belbin	
0.4	16/12/09	P. Flemons, B. Raymond, L. Belbin	Continuing revision of draft text
0.5	16/12/2009	P. Flemons, B. Raymond, L. Belbin	Fill in missing sections, add to refs and glossary, reformat
0.6	18/12/2009	P. Flemons, B. Raymond, L. Belbin	
0.7	18/12/2009	P. Flemons, B. Raymond, L. Belbin	
0.8	21/12/2009	P. Brenton	Re-format. Replaced Diagram 2 with updated version.
1.0	22/12/2009	L. Belbin	Final review and add group photo.
1.1	26/1/2010	P. Flemons	Integrate delegate responses
2.0	2/2/2010	L. Belbin	Final update and check
3.0	19/2/2010	L. Belbin, P. Dunstan	Piers submitted RAD details for Table 1 and Section 5.2.3

Table of Contents

1 Attendees

Person	Institution	Interests / Workshop Expectations
Jane Elith	University of Melbourne	Modelling distributions of species – identifying useful methods and making them available to end-users with sufficient info for understanding.
Leon Barmuta	University of Tasmania	
Simon Ferrier	CSIRO Entomology	ALA to incorporate large amounts of collections data (e.g., invertebrates) and including data & tools for environmental management Long term vision: - spatial tools to be readily available for use by anyone in the world with data conforming to standards.
Glenn De'ath	Australian Institute of Marine Science	
Dan Faith	Australian Museum	Keen to see aggregation of ideas into a useful form.
Jeff Tranter	ERIN	ALA to be a robust reliable clean definitive source of data
Jeremy VanDerWal	Centre for Tropical Biodiversity & Climate Change Research, Townsville	Interests: - Algorithms for modelling of spatial patterns of distribution & abundance – past present and future, and linking to genetic information. Also, algorithms associated with climate modelling and climate change. Workshop outcome: - To see a way forward on how to link up with terrestrial biodiversity network.
Michael Bode	University of Melbourne AEDA	Interests: - Prioritisation of social, political, economic, institutional aspects of biodiversity
Piers Dunstan	CSIRO, Marine & Atmospheric CERF Marine Biodiversity Hub	Interests: - Developing methods for prediction of multispecies distribution – mean estimate + uncertainty estimate. Integration of methods for including uncertainty in mapping distribution and incorporating uncertainty into decision-making processes. Workshop outcome: - Contribution of methods developed for multi-species analysis to date
Paul Flemons	Australian Museum	Interests: - Data management and biodiversity informatics, building web-based data access applications. Workshop outcome: - Contribute to and document workshop discussions and outcomes.
Ben Raymond	Australian Antarctic	Interests: - Data visualisation and exploratory

	Division	techniques, ecosystem modelling, integration and syntheses of data. Workshop outcome: - Contribution to establishment and ongoing development of geospatial toolkit
Lee Belbin	ALA Team Leader	Interests: - Value add ALA data, sustainability of the ALA and related work Workshop outcome: - Effective tools to assist in the development of policies for Australia's living resources and Australian scientific community.
Peter Brenton	ALA Analyst	Workshop outcome: - Documenting workshop discussions and outcomes,
Apologies		
Andrew Lowe		

Back: Jeremy VanDerWal, Glenn De'ath, Leon Barmuta, Simon Ferrier, Ben Raymond, Jeff Tranter, Dan Faith.

Front: Jan Elith, Piers Dunstan, Paul Flemons, Lee Belbin.

Absent: Michael Bode

Photo: Peter Brenton.

2 Workshop Aims

1. To identify analytical methods for addressing the following high-level use cases:
    a. Estimating the spatial distribution of biodiversity
    b. Identifying differences in biodiversity over space and time
    c. Prioritizing management actions based in part on biodiversity estimates and scenario analyses
    d. Identifying gaps in biodiversity information relating to spatial, temporal, taxonomic and environmental factors.
2. To identify the most appropriate methods to address the use cases. These methods must be
    a. Widely used and tested
    b. Accepted as State-of-the-Art.
    c. Robust
    d. Suit actual or anticipated ALA data (e.g. presence-only)
    e. Computationally tractable
    f. Able to be implemented cost-effectively
    g. Modular and extensible and therefore easy to build on and maintain
3. To identify the most effective option for implementation
    a. Integrated into the spatial portal
    b. Download data for desktop or mainframe analysis
    c. Or a hybrid solution
4. To address for accepted methods the following
    a. Input data
    b. Parameters
    c. Procedures and limitations
    d. Outputs
5. What strategy and resources would be required (people, time, money…) for the selected methods
    a. To build the applications
    b. To sustainably manage the applications

3 Audience for the Spatial Analysis Toolkit

The anticipated audience for the toolkit was the scientific community as advisers to policy makers on environmental management issues.

NOTE: Technical terms and acronyms are defined in the Glossary.

Prior to the workshop, the ALA conducted a series of interviews, questionnaires, and user needs analyses intended to document the requirements of the geospatial toolkit from likely users. The requirements were distilled down to four broad use cases (Section 2.1, above).

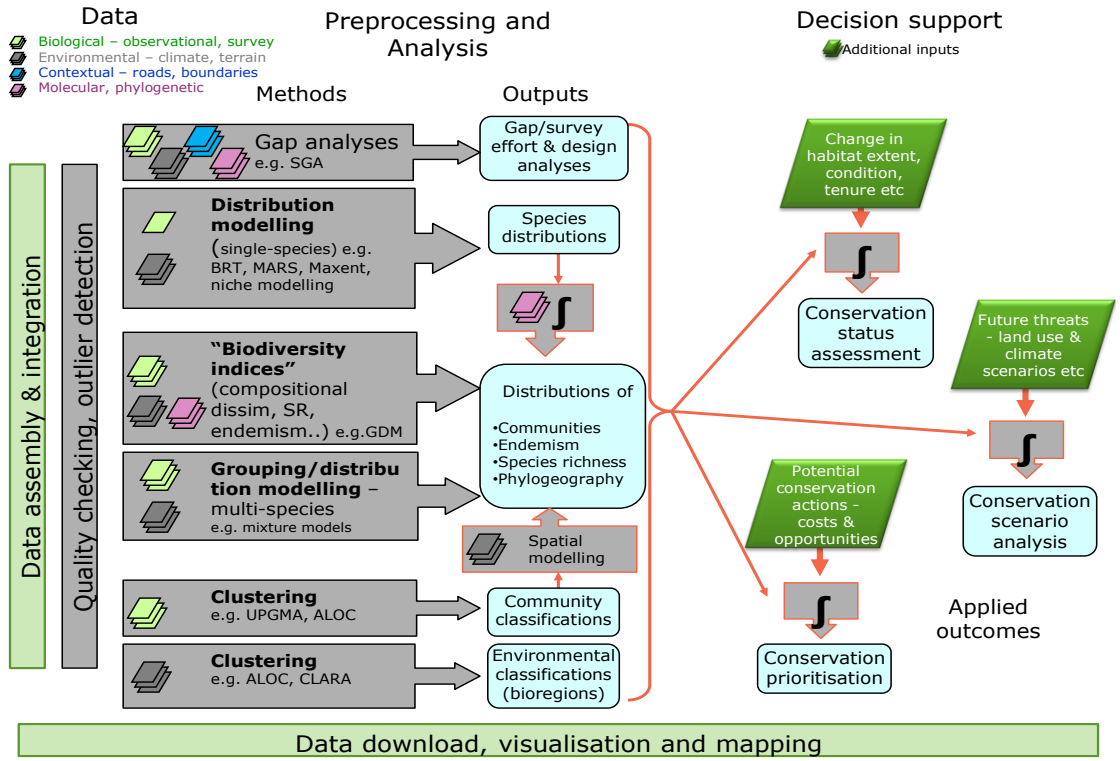
4 Workshop Strategy

Two approaches were considered for structuring discussion and exploration of analytical methods-

1. **User needs** - look at each use case and assess appropriate methods to solve it (noting that one methodology may address a range of use cases)
2. **Methodological** – identify relevant methods that may cut across a number of the identified use cases and proceed to explore those.

As a suite of methods were identified prior to the workshop that could address a range of priority use-cases, a methodological approach was agreed on. A first pass provided structure and the second pass filled in the details.

A round table discussion was used to explore Diagram 1 — components that the geospatial team identified as potential methods for the spatial analysis toolkit. This diagram was broken down by the group into pre-processing and basic analysis (left) and decision support (right).



The group identified components on the left of the diagram as more likely to be a priority for the ALA while the decision support components were more complex and likely to be difficult to implement and support within the ALA.

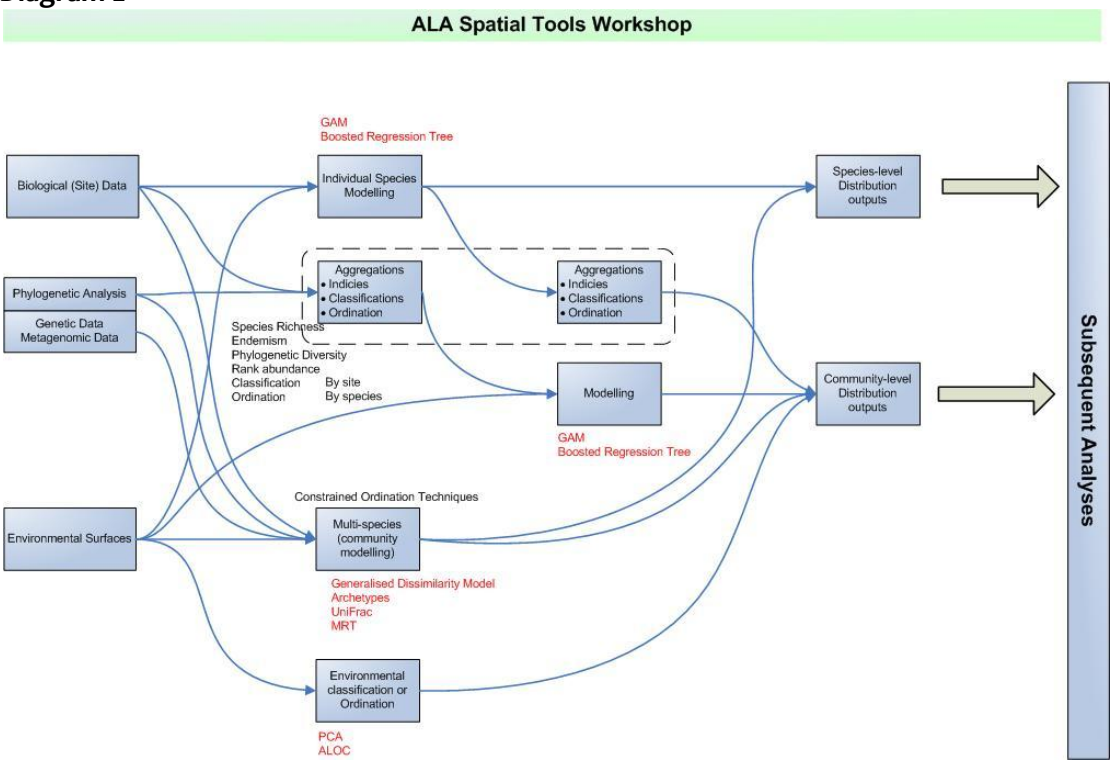
Diagram 2 (see below) developed by Simon Ferrier, provided a breakdown of the core methods into single species and community level analyses. Recommended algorithms and methods were identified within each box. Breakout groups were formed to address the following criteria within each box-

- Required inputs
- Recommended algorithms and their parameters
- Outputs produced
- Issues associated with all of the above including resourcing
- Workflows

Methods were classified into three groups

1. Individual species modelling including modelling of the outputs from data aggregation processes such as indices (e.g. species richness), classifications and ordinations.
2. Classifications and ordinations of
    a. Biological observation data
    b. Aggregations of individual species models
    c. Environmental data
3. Community-level and multispecies modelling

Diagram 2



Discussions raised a number of general issues about the spatial analysis toolkit.

What position the ALA should take on exposing users to the complexities associated with the tools? Should the ALA restrict potential scenarios to minimise the risk of solutions that may not be robust or should it provide broad functionality but require quality documentation and interfaces that would guide users? It was generally agreed that a balance was required.

- The ALA should focus on areas that addressed high-priority (frequently requested) use-cases. Data validation and cleaning was universally acknowledged as such a priority area
- While some domains were acknowledged as high-priority within the community, it may not be appropriate for the ALA to provide solutions. For example, it was agreed that the ALA should provide a limited number of IPCC climate surfaces for modelling, but should not provide area prioritization tools
- If there were multiple strategies for addressing priority use-cases, the most robust or simplest would be used
- Complex methods were required to address a range of needs, so effective guidance for users must be established
- Ideally, techniques should be able to accommodate presence-only data. Most of the current ALA data is presence-only but it is acknowledged that a broad range of survey data will be available through the ALA

5 Workshop Outcomes

5.1 Methods identified for implementation by ALA

Table 1: Use Cases: Estimating the spatial distribution of biodiversity; identifying differences in biodiversity over space and time. **Priority**: 1=implement, 2=implement if resources available, 3=would be nice, but...

Context	Method	Algorithms	Reasons for recommendation	Simple explanation of the method – inputs, outputs, value	Average Priority (1-3)
Single species modelling	Presence-only data	MaxEnt	Computationally fast, well programmed, good control options, good results, statistical rationale, well known and cited.	Inputs: presence-only or presence-absence data Outputs: individual species models Value: for presence-only or presence-absence data, these methods provide estimates of the probability of presence of the species in question, across the region of interest (not just at the locations at which species observations were actually made)	1
	Presence-absence data	Generalised additive models (GAMs)	Well understood, flexible, good statistical foundations.		1.75
		Boosted Regression Trees (BRT)	Ability to select variables, flexible, automatic fitting of interactions, good predictive accuracy, tree ensembles have increasing user base.		1.6
"Aggregation" methods	Classifications	UPGMA (Unweighted Pair Group Method with Arithmetic mean) and ALOC (non-hierarchical clustering)	Well understood, flexible (able to use arbitrary dissimilarity matrices). ALOC capable of operating on large data sets	Inputs: either biological observation data or predicted species distributions from models, and/or phylogenetic, genetic, or metagenomic data Outputs: indicators of ecological properties at community- or multi-species level	2

	Ordinations	(Non-metric) Multidimensional scaling	Well understood, flexible, fundamental method used in numerical ecology	Value: aggregation of single-species information to community level	1.9
	Indices	Various, including rarity and rank abundance distributions	Relatively simple descriptors of community-level biodiversity properties		2
Community-level and multispecies modelling		Rank Abundance Analysis	Modelling community structure (total abundance of all species, species richness and species evenness) using environmental gradients to analyse and predict Rank Abundance Distributions	Inputs: abundances of species found in a sample at different sites Outputs: Predictions of total abundance, species richness and species evenness with associated estimates of uncertainty	1.5
		Multivariate regression trees (MRT)	Yields interpretable explanations of ecological niches of communities	Inputs: presence-absence data of species across sites Outputs: various, including estimates of compositional dissimilarity between sites, phylogenetic diversity, groupings of species into communities, and the dependences of those communities on environmental predictors Value: direct modelling of community-level properties from biological observation and genetic/phylogenetic data	2.1
		Species archetypes (mixture modelling)	Able to handle data sets with rarely-observed species. Based on well-understood modelling methods. Simultaneously		2.2

			groups species and models group dependences on environmental predictors		
		Dissimilarity-based methods	A fundamental method in community ecology. Applicable to either species-observation data or phylogenetic data (e.g. phylogenetic diversity (PD) methods)		1.9
		Generalised dissimilarity modelling (GDM)	Applicable to a range of modelling tasks. Uses well understood methods. Performs well		1

5.2 Analysis Methods Requirements

5.2.1 Breakout Group 1: Single Species Modelling

Each point in the table is numbered for cross-referencing elsewhere in the report.

Analysis Methods	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
1. Single species modelling	2. Presen ce-only	3. Data quality is the major issue, including issues of biases (e.g. geographic clumping of survey effort), few records, and outliers or errors in the response variable	9. MaxEnt (strongly recommended to use Steven Phillips' version rather than the OpenModeller	10. Potentially suitable to be run as an online application 11. Geographic area of application (avoid extrapolation —	17. Scaled probability of presence of species across region of interest 18. Measures of uncertainty in	21. Delivering so as naive user can be guided in interpretation and model understanding 22. Download, saving

		4. Is the species in equilibrium with its environment, or is it (for example) still expanding in range? 5. Pre-processing of biological data indicating whether presence-only/presence-absence/abundance 6. Flexibility in what data fields to use in the analysis — e.g. may not want to use coordinates in the analysis 7. Based on data – warnings on number of records versus number of predictors 8. Guidance around categorical versus continuous data selection – ensuring that they are not incorrectly used	implementation)	predicting into unsampled environmental space) 12. Selection of background area 13. Automated detection of default parameter settings based on the data 14. Good user interface help , tutorials, information and links on using the various modelling methods, the implications of choosing different options etc, information to help interpret all components of the modelling process and the outputs 15. Batch processing – may be ok as long as good defaults set and good automated data cleaning and outlier removal. Also necessary to check that the response data are adequate to allow a model to be fitted (considering both number of data points and potential adequacy of	model predictions 19. Diagnostics from the models – e.g. measures of goodness of fit or predictive performance 20. Tools for interpretation of the model – how does the response vary for each environmental variable?	viewing options for outputs including model grids in various formats 23. Session management – saving all settings and input datasets – versioning etc. Logging of activities during session 24. Security management

				survey effort). User informed of caveats of these auto created models 16. Licensing of Phillips' package may be an issue as it isn't open source. Terms of use also preclude further distribution of the package		
	25. Presence-absence data	26. As above	27. Generalised Additive Modelling (GAM) 28. Boosted Regression Trees (BRT)	29. Potentially suitable to be run as online application 30. Geographic area of application 31. Algorithms available already in either R or C++ 32. Automated detection of default parameter settings based on the data. GAM and BRT not trained specifically on species modelling data (MaxEnt is) – need to setup defaults. Options to transform the data – and recommend, based on the data – what transforms or link functions to perform	35. As above Plus residual plots etc (not with MaxEnt)	36. As above

				33. Good user interface help – as per 14 (above) 34. Batch processing – as per 15 (above)		

5.2.2 Breakout Group 2: Classifications and ordinations

a. Of aggregations of individual species models
b. Environmental data

Each point in the table is numbered for cross-referencing elsewhere in the report.

Analysis Method	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
1. Abundance distributions	2. Abundance of every species within a community 3. Environmental predictor data	4. Requires abundances of all species within the community of interest (generally, all species present at each location) 5. Data limitations (quality and adequacy) – this method will only work with abundance data (not presence-absence or presence-only)	6. RAD Rank abundance distributions	7. Modelling of RADs as functions of environmental predictors is a relatively new technique	8. Estimates of richness, abundance, and evenness across the region of interest (with estimates of errors)	9. Abundance distribution methods are not as widely used as other methods discussed in this report, and so might require more effort in guiding users in their appropriate use
10. Rarity		11. Data limitations (quality and adequacy) 12. Need to define the whole community, to provide an appropriate context for assessing "rarity"				

Analysis Method	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
13. Classification of sites by species composition	14. A list of species at a number of locations (points or areas) - a sites by species matrix	15. Normalisation/weighting of areas and species 16. Requires presence-absence or abundance data	17. Hierarchical classification – Unweighted Pair Group Method using Arithmetic averaging (UPGMA)	18. Algorithm best limited to the number of sites meaningfully displayed on a dendrogram (~200?) 19. Selection of dissimilarity measure (Bray and Curtis OK as default)	20. Dendrogram (hierarchical diagram showing relationships between sites) 21. Set of groups (dendrogram cut at defined level) 22. Map of groups (points or areas)	23. Display of dendrogram difficult to interpret for more than ~ 200 sites 24. Determination of the number of groups
			25. Non-hierarchical classification (e.g. ALOC algorithm) using Bray and Curtis association measure	26. Handles large datasets (1 million+ sites) 27. Requires pre-selection of the number of groups. 28. It is possible to use non-hierarchical to reduce sites to site groups and then use hierarchical clustering to achieve the final number of groups. This combines the efficiency of non-hierarchical methods with the	30. Set of groups 31. Map of groups (points or areas) 32. Option: Intergroup distances matrix 33. Option: Table of site-centroid distances (could be used for 'representativeness')	34. The relationships between groups are not hierarchical 35. How do we map points (vs. areas)?

Analysis Method	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
				hierarchical information of the dendrogram 29. Selection of dissimilarity measure – as per 19 (above)		
36. Classification of sites by environmental variables (environmental domains)	37. A matrix of environmental variables sampled at sites	38. What environmental variables should be used for a particular domain definition? See 5.2.4.1.3	39. ALOC using Gower metric (range-standardized variables)	40. Handles large datasets (1 million+) 41. Requires pre-selection of the number of groups. 42. It is possible to use non-hierarchical to reduce sites to site groups and then use hierarchical clustering to achieve the final number of groups. This combines the efficiency of non-hierarchical methods with the hierarchical information of the dendrogram 43. Selection of dissimilarity measure – as per 19 (above)	44. Set of groups 45. Map of groups (points or areas) 46. Option: Intergroup distances matrix 47. Option: Table of site-centroid distances (could be used for 'representativeness')	48. The relationships between groups are not hierarchical
49. Classificatio	50. A species	51. Normalisation/weighting	53. UPGMA	54. Selection of	55. Dendrogram	58. Display of

Analysis Method	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
n of species by location	by sites matrix	of areas and species 52. Requires presence-absence or abundance data		dissimilarity measure, but probably Bray and Curtis default would suffice	(hierarchical diagram showing relationships between species) 56. Set of groups 57. Map of groups	dendrogram difficult for large number of species (more than approximately 200 species) 59. Mapping of species groups not simple (overlapping polygons)
60. Ordination of sites by species composition	61. A sites by species matrix	Normalisation/weighting of areas and species Requires presence -absence or abundance data	62. Non-metric Multidimensional Scaling (nMDS)	63. nMDS is computationally intensive but the limitation of the display probably outweighs the algorithm efficiency	64. Diagram of 2D or 3D coordinates of sites 65. Map of sites (using colour to indicate the position of the site within the ordination)	66. The ordination display may not be effective with a large number of sites (~200). 67. Interpretation of the display is specialized and ideally requires associated environmental data 68. How do we map points (vs. areas)?
69. Phylogeneti	70. Taxon	73. Integrity of the	75. Basic PD		76. Complementarity	78. The PD value is

Analysis Method	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
c Diversity (PD)	name 71. Phylogeny (specific or implied from taxonomic hierarchy) 72. Optional: (Probabilities of extinction)	phylogeny 74. Format of the phylogeny (standard?)			value (how much PD is lost if the taxon goes extinct?) 77. Expected PD loss	unbounded but could be standardised 79. Can get variation measure 80. Should display phylogeny link if it exists for taxon name selected. 81. PD30 or PD50??
82. Total PD of an area	83. As above, for user defined area	84. May limit to taxonomic group - Should apply species or taxon group filter	85. As above	86. As above	87. PD value (total PD lost)	88. As above
89. PD endemism (the evolutionary history lost for a site) analogous to species endemism	90. As above	91. May need to apply species or taxon group filter (e.g., mammals)	92. As above	93. As above	94. As above	95. As above

Analysis Method	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
96. PD complementarity	97. As above + additional set to obtain comparison	98. As above	99. As above	100. As above	101. As above	102. As above
103. PD dissimilarity 104. (covered under Dissimilarity methods in Section 5.2.3)						

5.2.3 Breakout Group 3: Community-level and multispecies modelling

Each point in the table is numbered for cross-referencing elsewhere in the report.

Analysis	Inputs	Input Issues	Algorithms	Algorithm Issues	Outputs	Output Issues
1. Multivariate regression trees (MRT)	2. Presence/absence or abundance data 3. Environmental data	4. Not applicable for presence-only data, but may be able to combine MRT with Ward et al.'s EM-algorithm to handle such data	5. Is currently implemented as R package 6. Relatively simple algorithm to re-implement if needed	7. Is a poor predictor. Improving predictions possible with e.g. bagging, but computationally expensive and sacrifices interpretability	8. Tree structure where leaf nodes describe a community type and the splits associated with it describe its environmental niche	
9. Species Archetypes This algorithm uses mixture models to simultaneously group species according to their environmental niches, and fit the responses of species groups to environmental	10. Presence-absence data 11. Environmental data 12. Possibly better suited than other algorithms to data with rare species	13. This method is new, although the underlying techniques (mixture modelling) are well established	14. Is currently implemented as an R package (C++)	15. Variable selection and interactions not well handled 16. Computationally expensive 17. Practical issues in fitting mixture models – can be difficult, but seem to work OK in this application	18. Group responses to environmental covariates, with standard errors 19. Probability of membership of each species to each group 20. Identification of an appropriate number of groups	

predictors					21. Residuals of species responses with respect too corresponding group responses	
22. Dissimilarity methods (for estimation of beta-diversity: species turnover across sites)	23. Distribution information of genotype, phylogeny, or species at sample sites		24. UNIFRAC implements dissimilarity-based methods for phylogenetic data, giving dissimilarities analogous to Bray-Curtis but based on phylogenetic branch lengths. UNIFRAC represents an important pathway for linking ALA work with genomics databases 25. Many algorithms for dissimilarity methods on species data (e.g. Bray-Curtis)	26. Existing software for UNIFRAC is free but not open source. Algorithms are relatively simple and much other software exists for dissimilarity calculations in general	27. Pairwise dissimilarities	28. Dissimilarities calculated by these methods are most likely to be used as inputs to further processing

29. Generalised Dissimilarity Modelling (GDM) Models compositional dissimilarity between sites as a non-linear function of the positions of those sites in geographic and environmental space.	30. Environmental layers 31. Presence-absence or abundance data		32. GDM currently implemented in C++, packaged both as a set of downloadable R functions , and as a stand-alone software application 33. Well tested and widely used	34. Potentially computationally expensive, but this is usually addressed effectively through automated subsampling procedures 35. Automated variable selection through Monte Carlo significance testing is computationally demanding for large datasets therefore usually relies on simple heuristic stepwise selection 36. Uncertainty in estimates not well handled, bootstrapping possible but slow 37. Interactions only handled at fairly simplistic level 38. Could benefit from	40. Core output: prediction of compositional dissimilarity between sites. This can be used for any downstream dissimilarity-based methods e.g. clustering, ordination 41. Secondary outputs: a series of transformed predictors (can be used for single-species distribution modelling via e.g. kernel regression)	

				better incorporation of geographic space into model 39. Techniques used to handle presence-only data could benefit from further refinement		
42. RAD Analysis: Analysis and prediction of rank abundance distributions	43. Abundances of ALL species found in samples at Sites. 44. Environmental Data	45. Requires counts of all species	46. The method is new, but has been published in Biometrics. Currently implemented in an R package	47. Current implementation uses a modification of a GLM framework. 48. Variable selection is via stepwise selection.	49. Prediction of Total abundance, Species Richness and Species Evenness + estimates of uncertainty.	50. Prediction with error can be computationally expensive for large numbers of points (i.e. 500000 +)

5.2.4 Workflows

5.2.4.1 *Common workflow components*

Many elements are common to more than one workflow. These are presented in some detail here, and referred to in each workflow.

5.2.4.1.1 Searching for data

1. User Search
    a. Select data from ALA cache
    b. Filter on any available fields, e.g. spatial accuracy, collector or date
    c. Upload own data
2. Returned data
    a. Records available – and range of metadata including
        i. Presence-only
        ii. Presence-absence
            1. From x number of surveys

5.2.4.1.2 Quality control

    b. Validation
        i. Check for outliers
            1. Visual (plots and maps)
            2. Automated stats output
    c. Summary statistics
        i. Box plot
        ii. Bar chart
    d. Automated assessment of data for informing analysis defaults
        i. Data structure
        ii. Data content

5.2.4.1.3 Selection of environmental layers

Environmental layers should be chosen appropriately for the specific details of the modelling task being undertaken. The number of layers should be kept to a minimum and chosen to be most relevant to the taxa or index being modelled. Ideally these will be direct measures of environmental factors that affect the distribution of the taxa. If direct measures are not available, indirect measures or proxies are commonly used.

Selecting the most appropriate layers for a particular taxon requires knowledge about the ecology of that taxon. Environmental factors that appear to control spatial distributions may be able to be discerned from an ordination of a sites/area by species matrix.

A workflow for layer selection might therefore look something like:

a. Scale/cell size selection – choose spatial and temporal scales that are appropriate for both the ecological scale of the taxon being modelled, and the application scale of the model output (e.g. a coarse-scale grid

would likely be inappropriate for a species that occurs over only a relatively small spatial range)

b. Review the list of layers available at the chosen scales, considering the most appropriate layers for taxa in question. For users who wish to model species that they are not knowledgeable about, it might be necessary to provide guidance on appropriate layers

At each step of the workflow, the metadata associated with each environmental layer should be available in order to guide the selection process. Metadata will typically include information on the processing steps used in the derivation of each layer, the spatial and temporal scales and extents of the data, the errors and uncertainties associated with the data, and other information that will guide this process.

The number of layers that a user selects is also a potential issue. A user might be tempted to include a large number of input layers, perhaps because:

- some methods (e.g. regression trees) perform variable selection as part of the model fitting process (that is, they select the variables actually included in the model from the pool chosen by the user), or
- the user does not have a good knowledge of the layers likely to be important for the taxa in question, and so it is tempting to throw in anything that sounds like it could be relevant.

The model fit will tend to improve with more input layers, but this can be due to over fitting — the model goes beyond fitting of the true relationships in the data and begins to fit spurious relationships (effectively, fitting the noise in the data). Techniques such as cross-validation can reduce, but not necessarily avoid, over fitting. Warning the user when they select too many predictor variables could therefore be a useful part of the interface.

5.2.4.1.4 Model assessment

The methods discussed in this report involve analysis and modelling techniques of varying complexities. The validity of any conclusions drawn from such models is critically dependent upon the validity of the model itself. Thus, an important part of many workflows is a model assessment step, wherein the model is checked to ensure that it is valid and the fitting process has operated correctly.

Model assessment is a diverse field and appropriate methods vary according to the analytical method being used, but include techniques such as the assessment of residuals (looking for deviations from assumptions, or biases that might indicate that the model has not fitted correctly). Assessment of predictive performance is a component of model assessment, but good predictive performance alone is not sufficient to conclude that a model is valid.

Investigating the relationships that have been fitted by a model can help not only with model assessment but potentially also with understanding the effects of environment on the taxon in question. Methods such as GAMs, BRTs and MaxEnt provide relatively direct methods for examining the fitted relationships between environmental predictors and

the response variable. Note: GAM and BRT require about the same level of effort – both require one line of code in R.

An issue in predictive modelling is the use of a previously fitted model to predict in situations for which the model is not appropriate. "Not appropriate" can take many forms – for example, any situation that violates the assumptions of the model. A common issue is the application of a model beyond the scope of the data that were used to train the model. For example, a model might have been fitted using training data covering a certain environmental envelope. Using that model to predict the response under environmental conditions outside of that envelope is termed extrapolation, and should be undertaken with extreme caution. This can be a valid exercise in some situations (e.g. environmental conditions only marginally different to the training data, with modelled relationships that vary relatively slowly with changes in environment).

Models that give confidence bounds (or error estimates) on their estimates can be advantageous in this situation because the errors will generally increase rapidly under extrapolation. The ALA spatial analysis tool kit should include functionality that assesses the data being predicted on, and warn the user if extrapolation is being undertaken. There are various techniques that can be used to assess the range of environmental conditions that are supported by a given training data set (and therefore beyond which the model cannot necessarily be applied): This strategy has been implemented in MaxEnt.

5.2.4.2 Single Species Modelling

1. User Search
2. Quality Control
3. Selection of area for prediction
4. Selection of environmental layers
   - Warning to user of selection of too many environmental predictor variables
5. Assess and deal with survey effort bias (presence only)
   i. Calculate survey effort from the data
      1. Target group background density (i.e. associated taxa or taxa captured with same/similar techniques)
   ii. Visual assessment
   iii. Define area from which background data can be chosen
      1. Manual
      2. Automated with ability to alter
   iv. Advice to user about using biological data from outside area of prediction
   v. Advice on how to deal with widespread species where there may be a number of disjunct populations (JT)
6. Choose process parameters
   a. Algorithm-specific parameters
   b. Scale
   c. Output
7. Choose output

a. Notification
b. Output type
    i. Print
    ii. Save
    iii. Export
    iv. Transfer to another process

5.2.4.3 Classifications and ordinations

These methods will be difficult to apply to presence-only data where there are scattered observations of a taxonomically wide range of species. In this case, the best solution is to have probability surfaces of the species of interest and sample those probabilities at a set of points in an area of interest.

1. Hierarchical classification of sites by species composition (see UPGMA)
    a. Select survey sites within an area
    b. Select a species of interest
    c. Steps a+b produce a sites by species matrix with content either presence-absence, abundance/cover etc., or probabilities from modelled surfaces
    d. Run classification and optionally select the number of required groups
    e. Examine dendrogram to understand relationships between sites and groups, and optionally select an appropriate number of groups
    f. Portray sites/areas on a map showing each group with a unique pattern or colour
2. Non-hierarchical classification of sites by species composition (see ALOC)
    a. Select survey sites within an area
    b. Select a subset of species of interest
    c. Steps a+b produce a sites by species matrix with content either presence-absence, abundance/cover etc., or probabilities from modelled surfaces
    d. Select number of groups required and run classification
    e. Portray sites/areas on a map showing each group with a unique pattern or colour
3. Hierarchical classification of species by sites (see UPGMA)
    a. Select species of interest

b. Select survey sites within an area

c. Steps a+b produce a species by sites matrix with content either presence-absence, abundance/cover etc or probabilities from modelled surfaces

d. Run classification and optionally select the number of required (species) groups

e. Examine dendrogram and select an appropriate number of groups

f. Species groups will occur across a subset of the sites so mapping is not straight forward, unless one species group is mapped at a time (with a pattern or colour)

4. Ordination of sites by species composition (see NMDS)

a. Select survey sites within an area

b. Select species of interest

c. Steps a+b produce a sites by species matrix with content either presence-absence, abundance/cover etc or probabilities from modelled surfaces

d. Run ordination (3d assumed)

e. Examine ordination (3d) diagram to identify patterns and processes

5.2.4.4 Community-level and multispecies modelling

5.2.4.4.1 Multivariate regression trees (MRT)

1. Biological data selection and validation
2. Environmental data selection.
3. Choose process parameters

   MRT has relatively few parameters to choose. It can operate either from the raw biological (site) data or from a dissimilarity matrix. For the former, one needs to choose whether to use Euclidean or Manhattan measures of node impurity (Euclidean is default). Using a dissimilarity matrix allows the user to choose an arbitrary dissimilarity function that is suitable for the data (e.g. extended dissimilarities). Other MRT parameters include the minimum number of cases per node, and number of cross-validation groups – these generally have fairly sensible defaults.
4. Output

   MRT gives a tree that describes environmental niche of each leaf node, and assignment of each input datum to groups (leaf nodes). Would want to be able to display tree to user, as well as map of group assignments. MRTs are not good predictors, and so while it would be possible, it would probably not be wise to use this as a tool for subsequent prediction of community types across gridded environmental data.

5.2.4.4.2 Species archetypes

1. Biological data selection and validation
2. Environmental data selection
3. Choose process parameters:
    a) Model structure. The species archetypes algorithm uses generalised linear models to relate biological responses to environmental gradients. Each species group (archetype) has an associated GLM. These GLMs are binomial (for presence/absence data) and the form of the GLMs must be specified (e.g. archetype response is a function of all environmental covariates plus their squared terms)
    b) Number of groups (archetypes). If this is not known (and generally, it will not be known) then the algorithm can be run multiple times (with a range of number of groups) and the most appropriate selected using information-theoretic (model selection) criteria

4. Output
   Model diagnostics and summaries are important for this technique, since it relies on reasonably elaborate modelling techniques.
   A number of outputs can be derived from the model:
    a) Each archetype will have an associated probability of presence at each sample location, and so maps of the distributions of each archetype can be produced (along with estimates of uncertainty in the probabilities)
    b) The model gives probabilistic estimates of membership of each species to each archetype (that is, the probability that species *x* belongs to archetype *y*). From this information, one can derive various forms of community composition (i.e. which species are contained within each archetype)
    c) The fitted dependence of each archetype on the environmental covariates (with uncertainties)
    d) The change in model goodness-of-fit with different number of archetypes (from which one can select the most appropriate number of archetypes)

5.2.4.4.3 Generalised dissimilarity modelling

1. Biological data selection and validation
2. Environmental data selection
3. Choose process parameters:
    a. Dissimilarity function. GDM models the compositional biological dissimilarity between sites as a function of the sites' environmental differences. An appropriate dissimilarity function must therefore be chosen — for presence-absence data, the well-known Bray-Curtis index would be a common choice. Inclusion of phylogenetic information into the dissimilarity function is possible (e.g. Faith 1992, Ferrier et al. 2007) and has already been trialled by Dan Rosaeur as part of a current DEWHA-funded project by linking GDM software to the Biodiverse package.

b. Spline parameters. GDM uses I-splines to generate nonlinear transformations of the environmental predictors. This transformation is carried out such that the compositional dissimilarities between pairs of sites are linear with respect to the differences between the (transformed) environmental conditions at those sites. The parameters of the splines (number and positions of knots) can be varied to achieve better model performance
c. Model structure. GDM uses a generalised linear modelling framework, and therefore requires the selection of a link function and variance structure. Appropriate choices will depend on the dissimilarity metric used — a discussion of these issues and appropriate choices for Bray-Curtis dissimilarities are provided by Ferrier et al. (2007). Interactions between variables also need to be specified by the user
d. Model selection. It is possible to apply various strategies to select the predictors actually included in the model (i.e. a subset of those chosen in step 2), such as stepwise selection. Similar strategies can also be used to guide the choice of spline parameters (see 3b, above). These strategies may or may not be appropriate for inclusion in the ALA toolkit, depending on the implementation framework (web- or desktop-based, target audience, interface design considerations, etc)

4. Output

   A number of outputs can be derived from the model:

   a. The primary function of the method is to provide estimates of the compositional biological dissimilarity between sites, given their environmental data. Thus, this model can be used to drive other ecological analyses that operate from dissimilarity data, such as classification and ordination. Survey gap analyses (searching for unsurveyed sites that are likely to provide biological information that best complements the existing data) are also possible
   b. The model fitting procedure involves the computation of nonlinear transformations of the environmental predictors (3b, above). The transformed predictors can then be used as the basis of species distribution models using relatively simple regression techniques (see e.g. Elith et al. 2006, Ferrier et al. 2007) or potentially any other existing species modelling technique (e.g. MaxEnt).

5.2.5 Resources Required

Method	Software	Availability	Funding Required	Personnel available	Time Frame
Single Species Modelling Presence-only data	MaxEnt	• MaxEnt is freely available through web download but is not open source	• None required for software or algorithm development • Development costs will mostly be focused on data preparation and user help and auto assistance mechanisms	Jane and Jeremy: • for working with programmers to identify data preparation process, data validation, parameter selection. Review. Possible workshop followed by iterative review. Contract/payment conditions to be negotiated with ALA. • Jeff Tranter keen to assist and provide test subjects	• Determined by web interface development timeframe
Single Species Modelling Presence-absence data	GAM BRT	• Algorithms available in a range of implementations of R and compiled executables in languages such as C++. They can be implemented through R or directly through the executables	• None required for software or algorithm development • Development costs will mostly be focused on data preparation and, and user help and auto assistance mechanisms	Glenn and Jane – BRT and GAMs: • for working with programmers to identify data preparation process, data validation, parameter selection. Review. Possible workshop followed by iterative review. Contract/payment	

				conditions to be negotiated with ALA	
Rank abundance					
Rarity					
Classification of sites by species composition	Dissimilarity measure, UPGMA	Probably R libraries in the first instance and easily programmable if a priority/speed required	2-3 days programming	Toolkit programmers (*2)	2-3 days

Classification of species by location	Dissimilarity measure, UPGMA	Probably R libraries in the first instance and easily programmable if a priority/speed required	2-3 days programming	Toolkit programmers (*2)	Included in the above
Classification of sites by environmental variables	ALOC	Simple algorithm – program from scratch.	5 days programming	Toolkit programmers (*2)	5 days programming
Ordination of sites by species composition	Dissimilarity measure (from above), NMDS	Probably R libraries in the first instance and programmable directly from published algorithms	5 days programming	Toolkit programmers (*2)	5 days programming
Phylogenetic Diversity (PD)					
Total PD of an area					

PD endemism (the evolutionary history lost for a site) analogous to species endemism					
PD complementarity					
PD dissimilarity (covered under Unifrac)					
MRT A form of constrained cluster analysis. Recursively generates homogenous groups in biological data by splitting on environmental predictor variables					

Species archetypes. Algorithm uses mixture models to simultaneously group species according to their environmental niches, and fit the responses of species groups to environmental predictors.					
Dissimilarity methods (for estimation of beta-diversity: species turnover across sites)					
Generalised Dissimilarity Modelling Models compositional dissimilarity between sites as a non-linear function of the positions of those sites in geographic and environmental space.					

5.3 Decision Support

The ALA could potentially assist decision support processes by providing tools that help assess specific spatial configurations. Assessment would be made in terms of the biodiversity represented in scenarios and conservation planning configurations.

There was insufficient time to discuss Conservation Status Assessment, Conservation Scenario Analysis and Conservation Prioritisation. It was however identified that the ALA should avoid:

- scenario analyses in terms of climate change , not so much due to the uncertainty in future climate scenarios but to the high level of uncertainty in the impacts of climate changes on biota, whether at the single species or community level, and
- developing specialist applications aimed at conservation prioritisation such as Marxan. Specialist expertise and data put these methods outside the scope of the ALA.

The mechanism by which the ALA could best contribute in these areas was agreed to be by means of providing a report card on the performance of any scenario or prioritisation configuration. The ALA could provide a set of tools that enable a user to upload in a standard format (e.g. shapefile) a scenario that could then be assessed against a range of biodiversity benchmark datasets or indicators. These report cards could then be compared, to ascertain which scenario may be preferred from a biodiversity perspective.

5.4 Issues Common across Methods

5.4.1 Biological Data

- Data Validation and cleaning :
    - clear summary of input data — e.g. visual presentation of different data characteristics through selectable/customisable symbology. Visual presentation of spatial reliability — e.g. depicting radius of uncertainty in location of data points.
    - understanding (provided by dataset and record level metadata)
        - survey effort
        - accuracy of recorded point locations
        - biases (e.g. spatial clumping of survey effort)
    - checking and correcting (annotation of original or correction of own set/local set) data for
        - taxonomic errors
        - georeferencing errors
    - pre-processing of biological data indicating whether presence-only, presence-absence or abundance
    - generation of pseudo-absences from presence-only data for methods that require presence-absence data.
- Ability to upload own data and an ability to apply validation and cleaning to uploaded data. Provide user with details on permitted formats - allow for range of formats in upload – feedback any format errors

5.4.2 Method

- Focus on doing small number of methods well and with adequate depth rather than doing a lot of methods superficially
- User understanding – combination of site documentation, links, video demos and GUI based mechanisms (e.g. pop-ups when poor option chosen) for:
  - guiding user in choosing most appropriate analysis options
  - guidance on limitations and uncertainties
  - Examples include
    - http://biodiversityinformatics.amnh.org/index.php?section=rsr_tools
    - http://biodiversityinformatics.amnh.org/index.php?section_id=7
    - http://biodiversityinformatics.amnh.org/files/SpeciesDistModelingSYN_1-16-08.pdf
- Pseudo-absence data – how should it be dealt with in relation to different methods? A tutorial on the implication of presence-only data would be useful

5.4.3 Environmental data

- Scaling environmental data to match input biological data. How do we decide what resolution of environmental data best fits the biological data? We may not have a choice but where possible should provide advice and warnings.
- Ability to upload or access non-gridded environmental data (e.g. point based marine, terrestrial or stream based surveys), explore/mine it and where appropriate extrapolate it to surfaces e.g. kriging
- Choosing /selecting most appropriate/relevant variables (Amazon approach to selecting environmental variables)
- Quality and provenance of environmental layers (metadata)
- Cross correlation of environmental variables - be good to be able to explore interactions between variables

5.4.4 System Functionality

5.4.4.1 *Online versus offline processing*

The tools described in this report could be implemented in a number of different ways. Should we provide online analytical tools within the ALA web portal or desktop-based tools that can are run by the user on their own machine? Possibly implement some tools that are simpler/ easier/robust to implement (eg MaxEnt and some of the cleaning tools) as online tools. The majority might be best used as desktop tools but ALA could provide good documentation, tutorials, help, examples and wikis for users.

The methods discussed here require a diverse array of supporting functions, from data reading and writing to display and graphing of outputs, and chaining of tools together to perform more complex analyses.

An online implementation will allow users to get up and running (using the tools) quickly but will require far greater resources and likely provide fewer options than a desktop-based software suite.

An online toolkit will have more direct access to the ALA's data holdings. Downloads of data to user-preferred applications may be prohibitive and may also be subject to licensing limitations.

5.4.4.2 *Batch processing*

The ALA could potentially provide pre-generated distribution models for some high-priority taxa and/or taxonomic groups. This issue has been discussed prior to, but not at the workshop and is therefore still under consideration. There was a recommendation from the ALA Team Leader's meeting (December 10, 2009) that we should consider birds and/or indicator species as a priority. Availability of comprehensive data (observations as well as specimen based) remains important. Extensive metadata will be required to accompany any canned models.

The reasons for canned models might include:

- Providing users with models generated by species modelling experts. Generating a large number of models would be prohibitive on any expert/group but ready access to probability surfaces would have high value to the ALA community, particularly if comprehensive species data was available. These models should be rerun whenever new species data becomes available.

- Providing users with a first-cut model of a larger number of layers. These models might not necessarily be of expert quality, but should be better than naïve users might be expected to produce themselves using desktop tools. Such models would be useful as a rough indication of species distributions or as a model that could be improved using the interactive tools. These models will need to be rerun whenever new species data becomes available.

Such modelling, particularly if implemented using automatic or semi-automatic algorithms would need to include quality checking and model assessment steps. These would include removal of outliers and errors prior to modelling, checking that the input data are sufficient to support such modelling (number and distribution of data points) and checking of model diagnostics and performance measures. Maybe there is an opportunity to advance the methods based on the volume of ALA and what we should be aiming at?

5.4.4.3 *Model fitting and assessment*

Many of the modelling techniques discussed here have a number of parameters and algorithm choices that can drastically affect the modelling process. Given usage by those not familiar with these modelling techniques, the tools should advise on the positives and negatives and where possible make sensible selections of model parameters based on the data chosen for a particular analysis task. This also includes the appropriate treatment of input variables: for example, recognising categorical and continuous variables and treating each appropriately

(or warning the user if a particular method does not work with categorical variables, for example); providing appropriate guidance on the transformation of input variables (those that require transformation in order to meet the assumptions of the algorithm, or in order to be more ecologically relevant).

The user might also be provided with guidance on input data — for example, checking the number of chosen predictor variables against the number of data points and warning when overfitting is likely. Survey bias is likely to be a significant issue and while it is not yet clear how this should best be tackled, it is recognised that assessing survey bias is a necessary component of the toolkit.

See also:

- selection of predictors (section 5.2.4.1.3)
- model assessment (section 5.2.4.1.4)

5.4.4.4 Session management

- Security Management – secure login and access, based on user profiles, to datasets and tools
- Logging of all activities – all activities associated with a users sessions should be logged for later review and reuse in running models again
- Saving all settings and input datasets for reuse or exchange with other users – including -
    - Versioning of input environmental datasets
    - Search criteria used for obtaining biodiversity data
    - The version of the cache at the time of searching
    - All algorithm/software parameters for an analysis
    - Version of algorithm/software
- Need to be able to monitor tool usage to gauge the sort of uses that tools are being applied to.

6 Conclusion

A comprehensive User Needs Analysis guided the selection of groups of methods going into the workshop. The workshop identified a suite of algorithms/methods that could be effective as a Spatial Analysis Toolkit for the Atlas of Living Australia.

It is widely acknowledged that the ALA would address a broad range of user needs by integrating biological and 'environmental' data in the portal. A priority must be placed on the availability of such data and the provision of effective methods for tabulation and display. Such methods should be considered an important component in the Spatial Analysis Toolkit.

When details of the methods and implementation requirements have been fully established, priorities for implementation will be allocated on the following criteria:

- Addressing priority user needs/ use-cases,
- Value adding to the ALA data holdings,
- People, time and computing resources required,
- Complexity of the methods (documentation and interface requirements for an effective implementation), and
- Anticipated availability of suitable data or priority to obtain suitable data.

It is anticipated that advice will be required from delegates and other specialists prior to and during implementation of the selected methods. Workshop delegates will be engaged in the process. A representative group of clients will also be engaged for testing of all aspects of the ALA site, including the Spatial Analysis Toolkit.

7 Glossary

ALA	Atlas of Living Australia
ALOC	A non-hierarchical classification method suited to large volumes of data (see Belbin, 1987). ALOC has four phases. The dataset is scanned once to generate a set of seed objects. Phase two allocates each object to its closest seed. Phase three calculates centroids based on group composition. Phase four is iterative: For each pass over the data, each object is removed from its group (centroid re-calculated without object) and the object is then allocated to its nearest centroid. The method converges quickly to a solution.
BRT	Boosted Regression Trees. A BRT uses a large number of small regression trees and combines them to produce a more accurate overall model (Friedman 2002).
GAM	Generalised additive models. GAMs are regression models that use non-parametric smoothing functions to describe nonlinear relationships. (Hastie & Tibshirani 1990)
GDM	Generalised dissimilarity modelling (Ferrier *et al.* , 2007)
Genetic data	
GLM	Generalised Linear Model.
IPCC	International Panel on Climate Change
Metagenomic data	The study of metagenomes, genetic material recovered directly from environmental samples
MRT	Multivariate regression tree. This is a form of constrained cluster analysis (De'ath, 2002). Recursively generates homogenous groups in biological data by splitting on environmental predictor variables
NMDS	Nonmetric multidimensional scaling. An ordination technique that accepts a dissimilarity matrix and a selected number of dimensions (usually 3). NMDS produces random coordinates in 3d, calculates Euclidean distances between each pair of objects, and performs a non-metric (ordinal) regression between the input dissimilarity and distance matrix, moves the points to maximise the fit and iterates until no movement improves the fit.
PD complimentarity	
Phylogenetic data	Phylogenetics is the study of evolutionary relatedness among various groups of organisms (for example, species, populations), which is discovered through molecular sequencing data and morphological data matrices.
Phylogenetic Diversity (PD)	

RAD	Rank abundance distribution. The RAD provides a summary of the abundance, richness, and evenness of a species community. RADs can be modelled in terms of environmental predictors (Wilson 1991, Foster & Dunstan 2009)
TERN	Terrestrial Ecosystems Research Network
UNIFRAC	Software tools for the comparison of microbial communities using phylogenetic information. It takes as input a single phylogenetic tree that contains sequences derived from at least two different environmental samples and a file describing which sequences came from which sample.
UPGMA	Unweighted Pair Group Method using Arithmetic averaging (see Belbin et al. 1992). An effective hierarchical clustering method where the distance between a newly formed group and all other groups is defined as the average of the distances between objects within both groups.

8 References

Belbin, L. (1987). The use of non-hierarchical allocation methods for clustering large sets of data. Australian Computer Journal, **19**, 32-41.

Belbin, L., Faith, D.P. and Milligan, G.M. (1992). A comparison of two approaches to beta flexible clustering. Multivariate Behavioural Research, **27**, 417-433.

De'ath, Glenn (2002). Multivariate regression trees: a new technique for modelling species – environment relationships. Ecology, **83**(4), 1105-1117.

De'ath, G. (2007) Boosted trees for ecological modeling and prediction. *Ecology,* **88,** 243-251.

Elith, Jane *et al.* (2006). Novel methods improve prediction of species' distributions from occurrence data. Ecography **29**, 129-151.

Elith, J., Leathwick, J. R. & Hastie, T. (2008) A working guide to boosted regression trees. *Journal of Animal Ecology,* **77,** 802-813.

Ferrier, S. Manion, G., Elith, J. and Richardson, K. (2007). Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions **13**:252–264

Friedman, J. H. (2002) Stochastic gradient boosting. Comput. Stat. Data Anal. 38: 367–378

Foster, S.D. and Dunstan, P.K. (2009) The analysis of biodiversity using rank abundance distributions. Biometrics doi: 10.1111/j.1541-0420.2009.01263.x

Hastie, T.J. and Tibshirani, R.J. (1990) Generalized additive models. Chapman & Hall/CRC. ISBN 9780412343902

Wilson, J. B. (1991). Methods for fitting dominance/diversity curves. Journal of Vegetation Science **2**:35–46