



ASRIS: Continental-scale soil property predictions from point data

Brent Henderson
CSIRO Mathematical and Information Sciences

Elisabeth Bui, Chris Moran, David Simon and Paul Carlile
CSIRO Land and Water

CSIRO Land and Water, Canberra
Technical Report 28/01, November 2001

Copyright

© 2002 CSIRO Land and Water.

To the extent permitted by law, all rights are reserved and no part of this publication covered by copyright may be reproduced or copied in any form or by any means except with the written permission of CSIRO Land and Water.

Important Disclaimer

To the extent permitted by law, CSIRO Land and Water (including its employees and consultants) excludes all liability to any person for any consequences, including but not limited to all losses, damages, costs, expenses and any other compensation, arising directly or indirectly from using this publication (in part or in whole) and any information or material contained in it.

ASRIS: Continental-scale soil property predictions from point data

Brent Henderson[†], Elisabeth Bui, Chris Moran, David Simon and Paul Carlile

CSIRO Land and Water

[†] CSIRO Mathematical and Information Sciences

CSIRO Land and Water Technical Report 28/01

November 19, 2001

Abstract

The Australian Soil Resources Information System (ASRIS) has compiled a national database of existing soil property information. This document details the point-modelling of soil properties from the ASRIS database. Continental soil property models are constructed from this database using decision trees. These relate the soil property to the environment through a suite of environmental predictor variables at the locations where measurements are observed. These decision tree models are then used to extend predictions, from known points to the entire continental extent, by applying the rules derived to the exhaustively available environmental variables.

The models developed have good to fair predictive ability. While they explain the broad changes in property across the ASRIS extent, their overall reliability varies spatially and depends on the property. In general, topsoil models are stronger than subsoil models. There is however a large amount of unexplained variation in all models. This is to a large degree expected given the heterogeneous nature of the database.

Certainty surfaces are provided for all predictions and are described here. They represent a combination of local point model performance, point density and environmental representativeness. The last two components emphasise the degree of extrapolation in extending predictions from the point data to the entire extent.

CONTENTS

1	Introduction	4
2	Data preparation	6
3	Point modelling strategy	8
3.1	Motivation	8
3.2	Environmental predictors	9
3.3	Statistical models	11
3.4	Variable selection	14
3.5	Model validation and assessment	15
3.6	Spatial dependence	16
3.7	Spatial implementation	17
3.8	Quantifying model certainty	17
3.9	Alternative strategies	20
3.10	Sampling issues and model biases.	21
3.11	Model performance and future directions	22
4	pH	24
4.1	Layer 1 pH in CaCl ₂	24
4.2	Incorporating data from Theme 5 Project 4D	36
4.3	Layer 2 pH in CaCl ₂	41
5	Organic carbon	48
5.1	Layer 1 organic carbon	48
5.2	Layer 2 organic carbon	55
6	Nitrogen	62
6.1	Layer 1 total nitrogen	62
7	Phosphorus	64
7.1	Layer 1 total phosphorus	64
7.2	Layer 1 extractable phosphorus	71
8	Clay	75
8.1	Layer 1 clay	75
8.2	Layer 2 clay	81
9	Texture	87

9.1 Layer 1 texture	87
9.2 Layer 2 texture	91
10 Thickness	94
10.1 Layer 1 thickness	94
10.2 Layer 2 thickness	99
11 Appendix: pH in water to pH in CaCl₂: A calibration equation	103
References	106

1 INTRODUCTION

The Australian Soil Resources Information System (ASRIS) has compiled a national database of existing soil and land resources data. This document details the point-modelling of soil properties from the compiled database.

Soil properties are modelled at a resolution of 250 metres for those river basins incorporating the intensively used agricultural areas. This extent is illustrated by the shading in Figure 1. It is broken down into 18 component regions with abbreviated region names as assigned. These jointly serve as separate modelling and assessment regions throughout this document. The ASRIS extent is large. At the 250 metre resolution there are over 43 million pixels at which predictions are required. Predictions are however aggregated to approximately a 1.1 kilometre (0.01°) resolution for presentation. Where the soil data do not support modelling for the entire extent a reduced extent may be used.

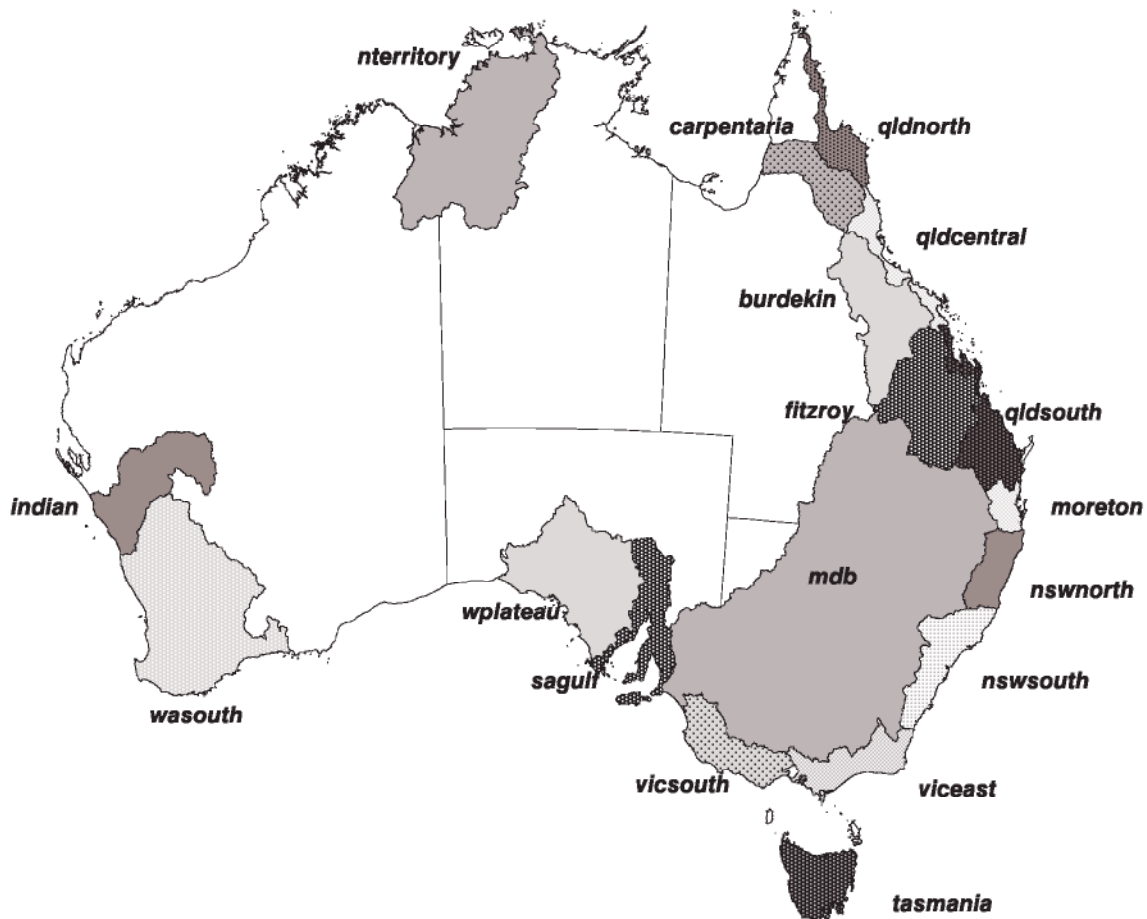


Figure 1: ASRIS extent and regions.

Those soil attributes considered are listed in the Table 1.

<i>Property</i>	<i>Layer 1</i>	<i>Layer 2</i>
pH	•	•
Organic carbon	•	•
Total phosphorus	•	
Extractable phosphorus	•	
Nitrogen (via C:N relationship)	•	
Clay	•	•
Field texture	•	•
Thickness	•	•

Table 1: List of properties modelled.

Layers 1 and 2 may be considered the topsoil and subsoil respectively. More specifically, layer 1 refers to the first A horizon and layer 2 the first B horizon recorded. If no horizon designators are available the first 30 centimetres are deemed the first A and thus layer 1.

The format for this document is as follows. Section 2 describes the data available and the preparatory steps taken. In Section 3 the general point modelling strategy adopted is outlined in detail. This includes discussion on the motivation for the approach, the modelling techniques selected, the predictor variables used and how the models fitted are validated. The notion of spatial dependence, the quantification of certainty in the predictions made and possible alternative modelling strategies are also visited. Sections 4 through 10 then consider each soil property individually by detailing property-specific data preparation and summaries. Model performance and certainty are also discussed.

2 DATA PREPARATION

After the initial quality control stage 135490 points were available in the ASRIS database for soil property modelling, albeit of varying completeness for individual properties. These break down as follows across the states/CSIRO.

NSW	QLD	SA	TAS	VIC	WA	CSIRO	Total
17885	48998	22085	3754	2674	38071	2023	135490

Table 2: Observation counts by State/CSIRO.

A small number of these observations fall outside the ASRIS extent. Moreover, the state tag is actually a pseudo-state tag because it is derived from agency codes which occasionally cross state boundaries. Examining the agency codes themselves gives the counts in Table 3.

agency count	102	199	202	211	212	290	299	300	301	311
	17428	457	1095	273	975	1	330	14270	5826	152
agency count	312	313	314	315	316	318	399	401	402	499
	1060	1085	4161	1014	5457	13779	2194	19960	663	1462
agency count	501	555	599	601	699	701	702	799	898	899
	36631	749	691	3482	272	62	436	92	7	1426

Table 3: Observation counts by agency.

Table 4 shows the break down across the 18 regions making up the ASRIS extent in Figure 1.

region count	nterritory	carpentaria	qldnorth	qldcentral	qldsouth	moreton
	705	226	647	11529	12103	6451
region count	burdekin	fitzroy	mdb	nswnorth	nswsouth	viceast
	8318	6577	17365	2593	6810	1042
region count	vicwest	tasmania	sagulf	wplateau	wasouth	indian
	3188	3605	12470	1265	33641	2457

Table 4: Observation counts by region.

Further data preparation involved conversion to common units/scales, investigating method differences and calibrating comparable methods where possible, interrogating unusual observations and deleting obvious outliers. This was done for each soil property separately and is considered in detail in the Section specific to each property.

The date at which the samples were taken or analysed was not considered due to the relative incompleteness of those fields in the database.

Multiple readings at the same location were allowed. Given we could not make the distinction between time series at the same point and observations that were taken at different locations but were actually tagged with the same point, it was decided to allow both.

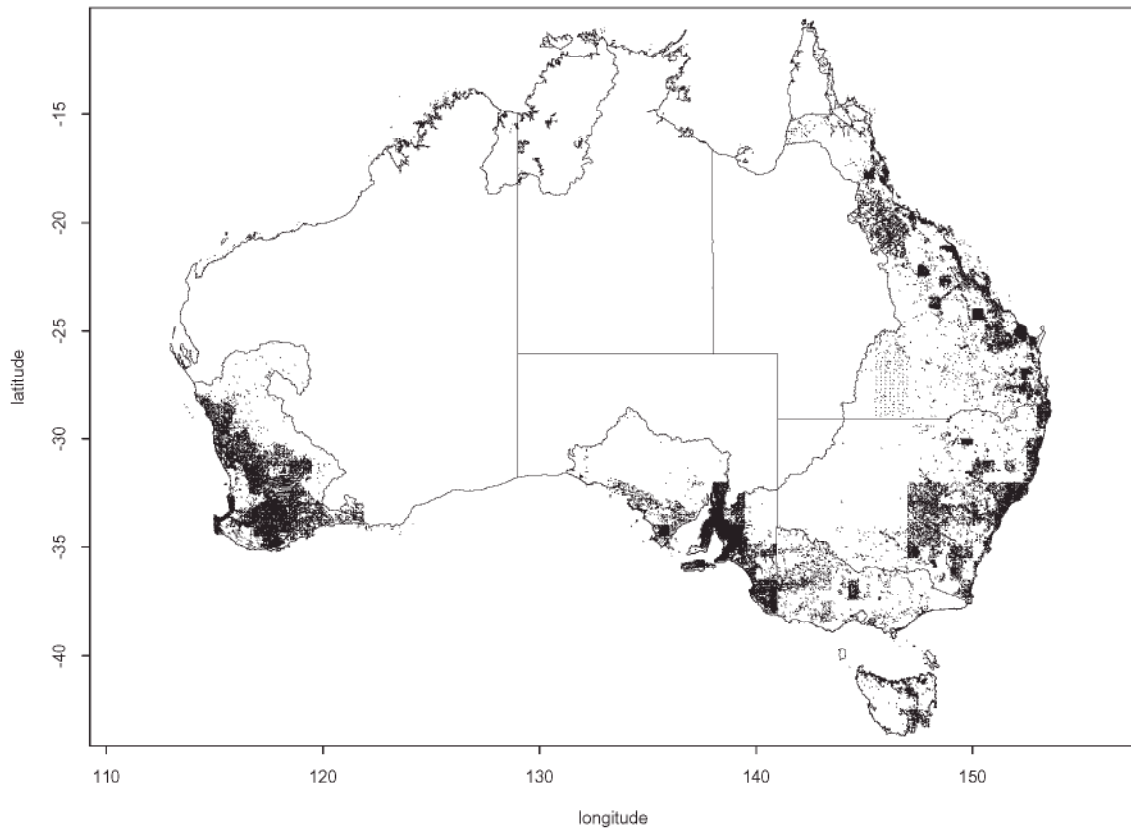


Figure 2: Locations of all ASRIS points

The locations of those points within the ASRIS extent are given in the Figure 2. It is evident that there are both areas of high intensity, where intensive local surveys have been undertaken, and sparse areas, where only exploratory surveys have been commissioned. This is to be expected given that the ASRIS database represents a collection of existing data sets from the various State and Territory natural resource management agencies and CSIRO Land and Water. At the agency level these data come from many sources. Some are derived from model-based or design-based surveys. Some are collected for specific purposes or intents and are thus often chosen in a subjective manner by appealing to experience and knowledge of landscape function or features so as to best answer the question of interest. Other samples are obviously dictated by convenience as they fall along roads or are otherwise more readily accessible.

This heterogeneity of sources and sampling techniques has some major implications for the modelling and validation, which typically perform best under homogeneous conditions. Issues stemming from this are considered in detail in the next section.

3 POINT MODELLING STRATEGY

3.1 Motivation

The *soil factor equation* has widespread acceptance amongst the soil science community. It asserts that soil characteristics are determined by five soil-forming factors, namely: parent material, climate, organisms, relief and time. That is,

$$\text{soil} = f(\text{parent material, climate, organisms, relief, time}).$$

These five soil-forming factors might be collectively termed the environment for greater brevity. Hudson (1992) states that this soil factor equation, “is simply a general statement implying that soils are natural bodies that are distributed in a predictable way in response to a systematic interaction of environmental factors”. There is no description of how these soil-forming factors actually affect soil characteristics, only that they do.

The great lesson from the soil factor equation is that differences in soil characteristics can often be attributed to differences between these soil-forming factors. If the nature of the relationship between the soil and these environmental factors can be established, then that relationship provides a vehicle to infer soil properties from knowledge of the local environment. This provides a natural complement to extensive survey work which can be a very costly exercise for large areas.

We used existing point observations and related them to environmental predictor variables at those sites. This is demonstrated in Table 5 where the point information, which consists of a point identifier and a property of interest, are supplemented by a list of environmental variables available at that point location. Note that mss1 and mss3 represent two bands from a multi-spectral scanner and ASC denotes the Australian Soil Classification. These are described in more detail in Section 3.2.

Point Information		Environmental variables						
point ID	soil property e.g. pH	elevation	mss1	mss3	annual mean precipitation	high period radiation	ASC	...
1	6.8	350	81	92	849	2417	kandosol	...
2	5.1	127	67	72	544	2589	vertosol	...
...
n	8.3	191	125	142	656	2387	calcarosol	...

Table 5: Example of point data supplemented by environmental variables.

A statistical model that captures the essence of the soil factor equation can then be constructed to relate the property of interest to (some of) the environmental variables available.

If these environmental variables are exhaustively available throughout the region of interest the model can then be used to extend predictions of the soil property from the existing point data set to the entire region.

The success of this approach relies on the strength of the relationship between the property and environmental variables available. Where these relationships are strong the

model will also be strong and the associated region-wise predictions given more confidence. Moreover, the benefit relies on these environmental variables being less costly to obtain than commissioning the collection of soil profile data and using some surface interpolation or smoothing procedure.

This type of approach has been well investigated in the literature. Much of the early development of the soil factor equation can be attributed to Jenny & Leonard (1934) or Jenny (1941) who sought to validate the soil factor equation by relating changes in soil properties to changes in climate across a West-East transect in the United States. Since then there has been a large body of literature that relates soil properties to landscape characteristics (e.g. Walker et al., 1968; Speight, 1974; Kreznor et al., 1989). Computer technology has made environmental variables more readily available and facilitated a greater pursuit of quantitative relationships with soil properties. For example, Moore et al. (1993) investigates quantitative terrain attributes for soil property modelling and Gessler et al. (1995) and McKenzie & Ryan (1999) both investigate statistical models for soil properties such as organic carbon and solum depth using modern environmental predictors.

There are however several important features of the ASRIS database to be considered in taking this approach, namely:

1. The 250 metre resolution at which we are working is notably coarser than the resolution at which others have worked. Some environmental variables that are useful predictors at less than 25 metres may not be so at 250 metres. Moreover, at 250 metres the within pixel variation for some properties may constitute a considerable proportion of the total variation, given expectations of local soil variability.
2. The size/extent of the problem is large. There are over 43 million pixels at the 250 metre resolution. Typically this approach has been employed in small catchments, which also have the virtue of being considerably more homogeneous. It is well known that both the pattern and process of variation will vary with different physiographic domains, and indeed that soil properties themselves will exhibit different ranges of variation (e.g. Beckett & Webster, 1971). Our focus is however necessarily on the long range variation given our goal is to produce continental-scale soil property predictions.
3. The ASRIS data derive from many sources and are collected with a range of underlying intents, as compared to what is typically a single survey in similar studies. There is also a disparity in the sample sizes available. Gessler et al. (1995) and McKenzie & Ryan (1999) base soil property models on at most a few hundred points, whereas there are thousands of points available from the ASRIS database for this modelling. That said, the sampling intensity relative to the extent is in all likelihood worse for the ASRIS data than in other applications of this methodology. Moreover, the ASRIS data will be subject to a greater variation in this intensity, with some areas only sparsely sampled, while others intensively surveyed.

3.2 Environmental predictors

A large number of environmental variables were considered to offer a useful quantification of the environment and be potential predictors of soil properties.

A suite of variables was derived from the continental AUSLIG 9" digital elevation model (DEM) and the drainage network to capture the landscape features and function. These included:

- elevation
- deposition path length
- erosion path length
- relative elevation
- relief
- slope percent
- hill slope length
- slope position
- river distance
- ridge distance
- contributing area
- inverse contributing area
- transport power in
- transport power out

These are collectively considered as the *terrain attributes*.

19 climatic surfaces were available:

- annual mean temperature
- mean diurnal change
- isothermality
- temperature seasonality
- maximum temperature of warmest period
- minimum temperature of coldest period
- temperature annual range
- annual precipitation
- precipitation of wettest period
- precipitation of driest period
- precipitation seasonality
- annual mean radiation
- highest period radiation
- lowest period radiation
- radiation seasonality
- annual mean moisture index
- highest period moisture index
- lowest period moisture index
- moisture index seasonality

These surfaces were derived from the Queensland Department of Natural Resources & Mines(QDNRM) climate data using internally-generated software. They were nominally available at a resolution of 5km, though were bi-linearly interpolated and resampled to a 250m resolution. These are described in detail in the meta-data.

A number of other environmental variables were available. These included:

- **MSS:** Bands 1-4 of the LANDSAT multi-spectral scanner (MSS).

- **Lithology:** Two lithology coverages, namely a 1:250000 and a 1:2 million scale coverage, from the Atlas of Australian Soils (Northcote et al., 1968; Bureau of Rural Sciences, 2000).
- **Landuse:** A 1:1 million scale landuse coverage from the Bureau of Rural Sciences.
- **Australian Soil Classification (ASC):** A 1:2 million scale digitized version from the Atlas of Australian Soils (Northcote et al., 1968; Bureau of Rural Sciences, 2000).

Other predicted properties were used in some models. These were typically the polygon surfaces created by a principal profile form (PPF) look-up table (Carlile et al., 2001). Some prediction surfaces from other point models were also used for related properties, e.g. layer 1 organic carbon was used as a predictive surface in the layer 2 organic carbon model.

It was necessary to transform some predictor variables prior to modelling.

It should be noted that some of these predictor variables were initially available at different scales. For example, MSS was originally available at a nominal 80 metre resolution, the climate surfaces at a 5 kilometre resolution and the digitized ASC on a 1:2 million scale. This was not seen as a problem, and possibly even an advantage, given soil-landscape processes operate over a range of scales, and that consequently, variation in soil properties may be explained by different factors at short and long ranges.

3.3 Statistical models

The models required fell into two broad types; classification and regression. For categorical properties such as texture we needed a classification model which, given a list of environmental variables at a given site, would classify the site response into one of a number of classes. For quantitative variables such as pH we needed a regression model which, given the list of environmental variables, would predict a continuous site response.

There was a large number of potential modelling techniques. Moore et al. (1993) used multiple linear regression models to predict a number of soil properties (A-horizon depth, organic matter content, extractable phosphorus, pH, % sand, % silt). Gessler et al. (1995) used the more flexible framework of generalized linear models to develop models for horizon thickness, solum depth and the presence/absence of E horizon. Gessler (1996) extended this flexibility by using generalised additive models which allow smooth functions of predictors to be included in the linear component of the generalized linear model. McKenzie & Ryan (1999) used tree-based methods to model soil depth and total soil carbon.

A decision tree-based methodology, through the modelling tools *Cubist* and *C5.0.*, was adopted for the ASRIS point modelling.

Cubist

Regression models were built for continuous properties (such as pH), using the modelling package *Cubist* which is available from www.rulequest.com.

Cubist has origins in the machine learning literature as a predictive modelling tool. It is the commercial successor to a technique known as M5 (Quinlan, 1992; Quinlan, 1993b)

which builds model trees using a decision tree methodology. The approach taken is similar to regression trees in CART (Breiman et al., 1984) but builds linear models rather than values on the leaves. The regression models it constructs are thus piecewise linear models.

The *Cubist* model tree is however converted to a series of rules, each with an associated linear model, to facilitate ease of interpretation. This involves some simplification of the set of rules derived from following the path from the root node to each leaf. Each *Cubist* rule is of the form

if {conditions} then linear model.

For example,

If
 slope < a
 annual mean temperature > b
 lithology = (p, q)
 then
 Property = $c_1 * \text{elevation} + c_2 * \text{mss2} + \dots$

where a and b are numerical values and p and q denote categories of lithology and c_1 and c_2 are coefficients of the linear model.

If the predictor variables associated with an observation satisfy the set of conditions, the linear model is used to predict the response. It is possible that any one observation and its associated predictor variables may satisfy more than one rule set in which case the average of the predictions is taken as the overall prediction.

A smoothing process is adopted to compensate for the discontinuities that may occur between linear models at different leaves. This is described in detail in Quinlan (1992).

Cubist uses a recursive partitioning of the predictor variable space in an similar way to the regression tree methodology of CART (Breiman et al., 1984). Both methods take a divide-and-conquer strategy and seek to minimise the intra-subset variation at each node. The criteria used to split a node is however slightly different because where CART uses the variance, *Cubist* uses the standard deviation as a measure of error. The reduction in error as a result of splitting a node is given by

$$\Delta_{\text{error}} = \text{sd}(T) - \sum_i \frac{|T_i|}{|T|} \times \text{sd}(T_i),$$

where T denotes the training cases available at that node, T_i represent the subset of those cases with the i th outcome following a given split, and $||$ denotes the count. The standard deviation of the response values is calculated for T and each subset T_i . The Δ_{error} then represents the expected reduction in error as a result of that split. *Cubist* chooses the split so as to maximise the expected reduction in error rate across all potential splits.

The advantage of the condition set in each rule is that it enables interactions to be handled automatically by allowing different linear models to capture the local linearity in different parts of the predictor variable space. This can often lead to smaller trees and better prediction accuracy than regression trees (Quinlan, 1992; Uysal & Güvenir, 1999).

In general, other advantages of tree-based regression are that they have a natural way to handle missing values, can use continuous and categorical predictors, are robust to predictor specification and make very limited assumptions about the form of the regression

model. Moreover, for large data sets a strength is that they do have the potential to uncover relatively complex structure that may be difficult to detect with more conventional regression tools.

Possible disadvantages that need to be considered are that there is no preference for low-order interactions, that they often do not make efficient use of the data and as such may not be useful for small data sets, that they are non-backtracking in that once a partition is made there is no consideration of alternative choices in the sub-tree (the implication of this is that each split is optimal and not the entire tree) and that they are not as well developed in terms of statistical inference as generalised additive models for instance.

Computationally *Cubist* is very efficient at deriving the regression models and extending the predictions to the entire ASRIS extent. This was seen as an important quality given the magnitude of the prediction problem.

Additionally, *Cubist* incorporates a number of modern statistical features as options which may be useful in some applications. It can allow the rule-based models to be complemented by instance-based or nearest neighbour information, cross-validation trials can be readily implemented to assess performance and it can use a committee of models to make predictions analogously to the “tree averaging” methodology of bagging and boosting.

C5.0

Classification models were constructed for the categorical properties such as texture using the modelling package *C5.0* (Quinlan, 1993a). This is the successor to C4.5 and is commercially available from www.rulequest.com.

C5.0, like *Cubist*, has a machine learning heritage. It is a classification tool that adopts a decision tree methodology to construct a classifier. This uses a recursive partitioning strategy, analogous to CART (Breiman et al., 1984), that makes splits on the predictor variables so as to create a progressively more homogeneous data set.

These decision trees can be converted to a series of rules so as to be more readily understood. This involves some simplification of the full set of rules which would stem from considering the path from the root node to each leaf. Full details are given in Quinlan (1993a). The *C5.0* rules used are of form

If	slope < a
	annual mean temperature > b
	lithology = (p, q)
then	
	Property = category II

where a and b are numerical values and p and q denote categories of lithology.

The decision on the optimal split at a given node is made according to the *gain ratio criterion* which is by definition the ratio of the *gain* to the *split info*. The *gain* amounts to the change in entropy between the node and the weighted entropy across the sub-nodes stemming from the split. The *split info* is a modifying quantity used to avoid bias in favour of splits with many outcomes.

More specifically, the *gain* of a split X is given by

$$\text{gain}(X) = \text{info}(T) - \sum_{i=1}^n \frac{|T_i|}{|T|} \text{info}(T_i),$$

where T represents the set of training cases at the node, T_i the set of training cases at the i th sub-node following split X , n is the number of outcomes from split X and $||$ gives the count. $\text{info}(T)$ and $\text{info}(T_i)$ denote the average information of set T and T_i respectively (also known as the entropy), where

$$\text{info}(S) = - \sum_{j=1}^k \frac{\text{freq}(C_j, S)}{|T|} \log_2 \left(\frac{\text{freq}(C_j, S)}{|T|} \right)$$

for set S , C_j identifies the j th class and k is the number of classes. The *split info* of a split X is given by

$$\text{split info}(X) = - \sum_{i=1}^n \frac{|T_i|}{|T|} \log_2 \left(\frac{|T_i|}{|T|} \right)$$

and is large when there are no dominant groups in terms of counts. The *gain ratio* of a split X is then defined as

$$\text{gain ratio}(X) = \text{gain}(X) / \text{split info}(X).$$

This quantity measures the proportion of the information generated by the split that is beneficial to the classification. *C5.0* consequently selects the split so as to maximize the *gain ratio* and thus the information gained. A detailed description of the criterion and its motivation is given in Quinlan (1993a).

In general, advantages of *C5.0* decision tree methodology are similar to CART and to those listed for *Cubist*. Both have a natural way to handle missing values, can use continuous and categorical predictors, are robust to predictor specification and make very limited assumptions about the form of the classification model.

Possible disadvantages that need to be considered are also similar to *Cubist*. There is no preference for low-order interactions, that they often do not make efficient use of the data and as such may not be useful for small data sets, that they are non-backtracking in that once a partition is made there is no consideration of alternative choices in the sub-tree (the implication of this is that each split is optimal and not the entire tree).

C5.0 is widely recognised for its performance and speed. Lim et al. (2000), in a comparison of 33 classification methods over 32 different data sets, found that *C5.0* performed well, especially in terms of speed but did have a tendency to produce larger trees than some of the other decision tree methods.

The *C5.0* software allows differential misclassification costs, fuzzy classification, boosting and assessment of performance using cross-validation which may be additional advantages for some applications.

3.4 Variable selection

The environmental variables used in *Cubist* and *C5.0* were selected after considering utility across a number of investigative measures.

The individual predictive power of the environmental variables was investigated through examining correlations with the property of interest, and by considering the reduction in deviance from fitting additive models with a smooth function of each environmental variable as a predictor in turn. Those variables with largest reduction in deviances were deemed most useful.

A forward stepwise linear regression procedure was implemented to consider multivariate relationships and the most important predictors identified. A regression tree model was fitted in S-PLUS using the RPART (Therneau & Atkinson, 1997) routine in an attempt allow non-linearity in these multivariate relationships. The most important predictors to the RPART model were similarly identified.

For categorical predictors an ANOVA using the predictor as the sole explanatory variable was used to assess variable importance.

Cubist does not directly provide a measure of predictor importance. The set of rules can however be examined and the frequency of the individual predictors in the conditioning and the linear model used as a measure of the utility of the predictor.

Strong correlations between predictor variables over the 135490 points in ASRIS database were identified as possible sources of collinearity. This was of some concern in the linear model at each leaf in *Cubist*, but not of concern elsewhere or in *C5.0*, because collinearity is only a source of non-robustness in that it can generate alternative splits that perform almost as well.

The final decision on what environmental variables to include as predictors in the *Cubist* and *C5.0* models was made by subjectively pooling all assessment measures considered.

3.5 Model validation and assessment

All models were developed by 70:30 training to test data split. 70% of the observations were used to construct the model in the model development stage. 30% were held back in order to assess the performance of the model. In view of the unequal representation from the States/CSIRO this 70:30 split was maintained within each state by treating each State and CSIRO as strata.

Once the training model was deemed optimal it was assessed on the test data. A final model was then fitted using all the data with the same model form and options. The performance of the final model was then assessed by 10-fold cross validation. This involved randomly dividing the data into 10 partitions or folds. At each step, nine of these partitions were used to fit the model and the performance assessed on the remaining partition held back as the test data. This procedure was repeated for each partition sequentially. The performance, averaged over all 10 partitions held back, gave the cross-validated performance assessment.

The performance of models at all stages was assessed in terms of a number of key indicators. These included:

- N : the number of points used in the model.
- R^2 : an estimate of the percent of the overall variation in the property explained by the model. Also known as the coefficient of determination.
- *RMSE (Root Mean Square Error)*: This is an estimate of the standard deviation of the errors. A lower RMSE is associated with greater predictive ability. RMSE values

can not however be compared between different properties because they depend critically on the scale used.

For a validation data set (test data) of size m the RMSE is defined as

$$\text{RMSE} = \sqrt{\sum_{j=1}^m (y_j - \hat{y}_j)^2 / m},$$

where \hat{y}_j denotes the predicted value.

- *Correlation*: the strength of the linear association between the observed and the predicted values. The rank correlation calculates this on the respective ranks rather than the actual values.
- *Average error*: This is the average absolute difference between the observed and predicted values. That is

$$\text{average error} = \frac{1}{m} \sum_{j=1}^m |y_j - \hat{y}_j|$$

Lower average errors tell us that the predicted values are closer to the observed values more often. Average errors do however depend critically on the scale of the units used and so can not generally be compared between models. This is also known as the mean absolute deviation.

- *Relative error*: This is defined as the ratio of the average absolute error magnitude to the error magnitude that would result from predicting the mean value, i.e.

$$\text{relative error} = \frac{\frac{1}{m} \sum_{j=1}^m |y_j - \hat{y}_j|}{\frac{1}{m} \sum_{j=1}^m |y_j - \bar{y}|}$$

If there is little improvement on the mean, the environmental variables have little predictive capacity and the relative error is close to 1. Generally the smaller the relative error the more useful the model.

- *Classification rate*: For the categorical properties the overall probability of correct classification from the decision tree is reported. This classification rate is simply 1 – overall error rate. A category-specific classification rate may also be reported for each category which gives an indication of the accuracy with which observations from the category of interest are predicted.

3.6 Spatial dependence

While no direct account was made of the spatial structure, it was assumed that the spatial dependence was implicitly accounted for by the environmental predictors. Spatial structure did however persist in some model residuals. Possible reasons are:

- that there are unspecified or unidentified environmental predictors which explain variation in the soil properties

- that there are differing scales of resolution (e.g. the climatic predictors have a resolution of 5km which necessarily implies they have no ability to explain shorter range variation than 5km, i.e. they may explain the longer range signal but will have no predictive ability over ranges less than 5km.)
- that the models developed are at a continental-scale versus at a more local level. This means that those environmental predictor versus property relationships that persist over the entire continent are the most likely to be identified, possibly at the expense of more local relationships which capture the spatial dependence.

Residual spatial dependence was investigated through variogram analysis.

3.7 Spatial implementation

Cubist and *C5.0* models both can be described in terms of a series of rules. These rules were implemented through a combination of purpose written C++ code and GIS software to produce the digital maps. Both *Cubist* and *C5.0* do incorporate facilities with which to make predictions for unknown observations but it was more efficient to take the approach adopted for this spatial context.

Where models were derived on a transformed scale, e.g. $\log(\text{property})$ or $\sqrt{\text{property}}$, the predictions were initially made on that scale and then back-transformed to the natural scale. Note that a correction for bias was necessary in these instances. For example, if a model is derived for $\log(\text{property})$, predictions on the property-scale are given by $\exp(\hat{\mu} + \hat{\sigma}^2/2)$, where $\hat{\mu}$ is the prediction and $\hat{\sigma}^2$ the estimated residual variance on the log-scale, rather than simply $\exp(\hat{\mu})$ as might be expected. In the case of properties modelled on the square root scale, the bias-corrected back-transformation is given by $\hat{\mu}^2 + \hat{\sigma}^2$, where $\hat{\mu}$ is the prediction on the square root scale and $\hat{\sigma}^2$ is the estimated residual variance on that scale.

3.8 Quantifying model certainty

Three components were considered to contribute to the assessment of model certainty, namely the:

- point density
- environmental representativeness
- spatial model diagnostics

Point density

The point density was deemed important because it reflected the *physical proximity* of points in the ASRIS database for that property. This was seen to impact directly on certainty in predictions, because for areas where there was a higher point density, there was greater local representation and thus the expectation of a better account in the modelling. A smoothed representation of the raw point density at each pixel was produced by counting the number of points within a 15km radius.

Environmental representativeness

The aim of the environmental representativeness surface was to assess the *environmental proximity* to points in the ASRIS database for the property of interest. This was considered a valuable indicator of model certainty because areas whose survey intensity was low, may still be predicted with reasonable confidence if they were environmentally similar to a large number of observations in the database. The environmental representativeness surface thus sought to identify those environments that were well-represented in the ASRIS database.

The approach chosen to create this surface was to take a systematic grid sample of approximately 0.1% of the 250m pixels (some 44798 locations) from the ASRIS extent. At these sampled locations a set of p environmental predictor variables was selected that characterized the environment. For example, values for variables elevation, mss2, mean annual temperature, mean moisture index and relief might be sampled and attached to each location in the grid. The choice of the variables used was governed by predictor variable importance for the property of interest. For each location in the ASRIS grid sample, we calculated how many observations in the ASRIS database, for the property of interest, were within some environmental distance threshold. A large count indicated an environment that was well represented in the ASRIS database. A small number, on the other hand, identified an environment that was not well represented, and thus an area where there should be less certainty in the property predictions. The lower representativeness was thus considered to reflect an area where there was a higher degree of extrapolation.

The environmental distances were calculated by standardizing to normal scores those predictor variables selected in the ASRIS database. Then, for each location in the ASRIS grid sample, the value of every individual predictor variable was transformed separately into a normal score. This was achieved by first identifying the rank of the value within the ASRIS database for that variable, and then assigning an interpolated normal score. For each location in the ASRIS grid sample, the Euclidean distance of these standardized environmental variables to all points in the ASRIS database was calculated, and the number within a distance threshold τ recorded, i.e.

$$c_i = \sum_{j=1}^m I \left[\sum_{k=1}^p (ze_{ik} - za_{jk})^2 < \tau \right]$$

where c_i is the count for the i th point in the gridded sample, j indexes the m observations in the ASRIS database for the property of interest, ze_{ik} is the k th standardized environmental variable for the i th point in the ASRIS grid sample, za_{jk} is the k th standardized environmental variable for the j th observation in the ASRIS database and $I[\cdot]$ is an indicator function that is 1 if the condition is true and zero otherwise. This count c_i was then used directly as a measure of environmental representativeness within the ASRIS database for the points in the gridded sample. It was extended to the entire extent by interpolation to the 250 metre resolution.

The choice of distance threshold τ was arbitrary, though was supported by considering small percentiles from a χ_v^2 distribution, given that would be the approximate distribution of the Euclidean distance if v independent variables were used.

Spatial model diagnostics

Statistical models were fitted to the ASRIS extent in its entirety. In order to provide an understanding of performance spatially, model diagnostics such as the relative error or classification rate were initially tagged to 18 regions that exhaust the ASRIS extent in Figure 1. It was subsequently decided that smaller regions would be more insightful and the extent was decomposed into 58 regions shown in Figure 3. These regions were created by intersecting the Australian Water Resources Committee (AWRC) basins and the ASRIS extent, and aggregating those basins inside the extent and smaller than a minimum subjectively chosen basin size.

These 58 regions were then assigned local model performance statistics to help identify where the model performed well, and thus where greater confidence in the surface could be placed. The relative error was used here for the piecewise linear models. In defining the relative error, the local absolute error was divided by the average deviation from the local mean (i.e. \bar{y} is the regional mean not the global mean). For classification models the classification rate was used. All pixels in a region were assigned the same value. If fewer than 20 points were available for any of the 58 error regions, the relative error or classification rate was deemed to be estimated unreliably and assigned a value of 0.

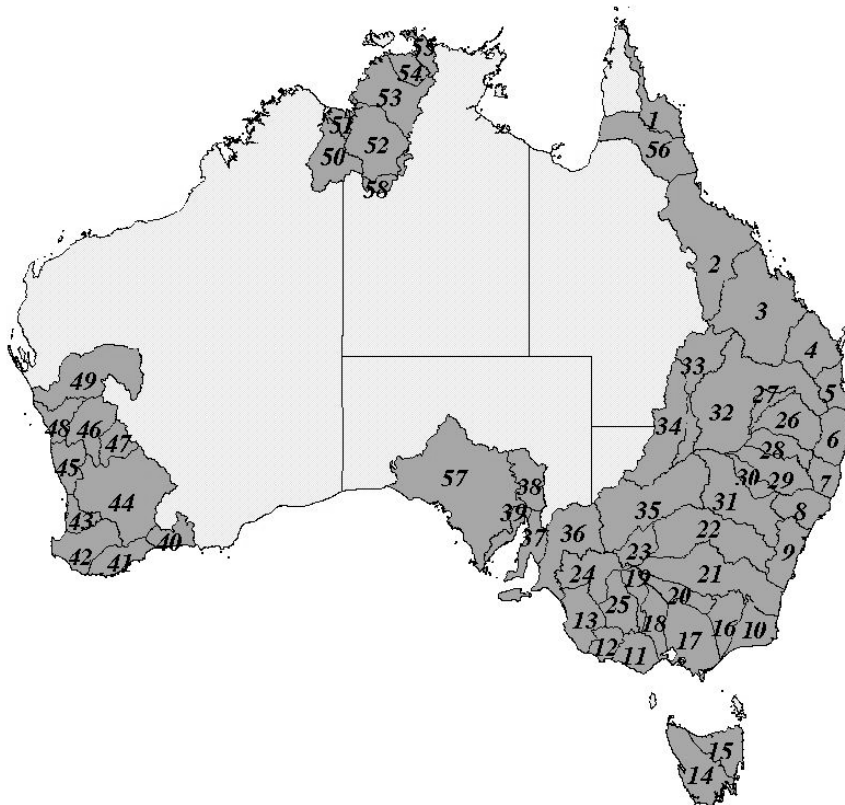


Figure 3: Regions used to assess model performance.

Model certainty

The final certainty figures presented combine all three components after standardisation to a 0–1 scale. This was achieved for the point density by clipping the count per 15km radius to 300, taking $\log_{10}(\text{count} + 1)$ and dividing by $\log_{10} 301$ to create a proportion between 0 and 1. A 0 therefore indicated no points in the 15km radius and thus no local support, while $2/3$ inferred 30 points, and 1 denotes 300 or more points, and thus very good local support.

The environmental representativeness was converted to a 0–1 scale by ranking the counts of environmentally similar points in the ASRIS data, and ascribing the rank as proportion of the 44798 points in the Australia-wide grid sample as an index of similarity; 0 being the least well represented and 1 being the most well represented locations.

The classification rates required no standardisation. The relative errors were however considered as $1 - \text{relative error}$, given the strength of support occurred in the opposite direction to the point density or environmental representativeness. Relative errors larger than 1 occurred in some regions when performance from the model was worse than the local mean. These were however clipped to a relative error of 1 before inclusion in the certainty surfaces.

The three component surfaces are represented on different bases. The point density is on the 250 metre pixel basis, the environmental representativeness on a grid sample basis and the spatial model diagnostics on a region basis. These were however converted to the common 250 metre pixel basis by interpolating the environmental representativeness grid sample and re-sampling the regional spatial model diagnostics to a 250 metre resolution. A weighted average of these three standardized components was then given as the model certainty surface, with higher values used to indicate greater model confidence. An example of how these components combine is given for layer 1 pH in Section 4.1.

3.9 Alternative strategies

The traditional adjunct to the soil survey is to somehow interpolate between or smooth the observed points so as to make predictions at unknown locations. This is the approach geostatistics takes through various forms of kriging. This type of approach is however most useful when the point observations are relatively dense and distributed throughout the region of interest. When there are regions where observations are particularly sparse, there is little local information with which to make predictions and thus little confidence in the predictions made.

Environmental and climatic predictors may be incorporated into the modelling through variants of kriging like regression kriging, kriging with an external drift or co-kriging (e.g. Odeh et al., 1994). This is however typically of most benefit when the environmental predictors are few in number and strongly correlated with the response of interest, neither of which were found to be true in this context.

Greater support for the point data might be derived by incorporating expert knowledge if it exists. Cook et al. (1996) uses a rule-based system to predict organic carbon where predictive attributes such as topography or rainfall are weighted according to prior knowledge of their association. Kiiveri & Caccetta (1998) incorporate this type of knowledge to salinity mapping via conditional probability networks.

While decision tree methods have been described here, other regression techniques to capture the soil factor equation were investigated. In particular, generalized additive

models (GAMs) were considered. In comparisons on test data sets, the performance was found to be very similar. The GAMs enabled a better modelling of mean-variance relationships, and were possibly more efficient in their parameter use, but were considerably computationally more demanding in making predictions. Given the size of the data sets available and the magnitude of the prediction problem, the tree methodology was adopted. Multivariate adaptive regression splines (MARS) (Friedman, 1991) may however represent an alternative to *Cubist*. While they were discounted because of their heavier computational burden and their reduced ability to cope with high dimensionality compared to *Cubist*, they represent an option that could be entertained in the future. MARS has the advantage over GAMs in that it can automatically account for interactions.

3.10 Sampling issues and model biases.

The ASRIS database is not a homogeneous quantity. There are many sources of variation in the database. Differences due to sampling regimes, laboratories, agencies, methods and techniques, the level of accuracy recorded, quality of the sample, measurement error, handling of non-response, and the date at which the sample is taken all contribute to the variation in the properties we see.

There is a potential for some of these factors to introduce biases into the soil property models because differences in some of these factors may be confounded with changes in property values. Corrections can be made for some factors to reduce this possibility, such as method calibrations (e.g. pH in water and CaCl_2). For the most part, however, the database is a compilation of existing data and there is limited scope to isolate or negate the effect of these factors on the derived soil property models.

One of the larger expected sources of model error here is due to the heterogeneity in the sampling regimes. de Gruitjer (2000) provides a good review of statistical methods in sampling for spatial inventory and monitoring of natural resources. de Gruitjer makes the distinction between three modes of point sampling.

- *convenience*: where those points selected are governed by being more readily accessible or measured, e.g. near road sides.
- *purposive*: where the points selected are chosen so as to address a particular question, e.g. monitor acidification
- *probability*: where the points are allocated at random locations.

Under probability sampling the selected points are necessarily representative because each point in the region has an equal chance of occurring (under simple random sampling). The danger of the first two modes, which make up the great majority of the ASRIS database, is that they introduce potentially sizable biases into the modelling because the data do not adequately represent all physical or environmental domains into which we are making predictions. Moreover, there is possibly even a preference for atypical or unrepresentative points in some instances. This might for example occur if an area is largely considered “known” and there appears greater value in selecting points so as to characterise those parts of the area which are more unusual.

The derived point models are internally validated here, either via a training/test split or via cross validation. This provides an assessment of performance against the data we have observed but does not infer how well the models might perform across the whole

extent because the dataset is necessarily a biased sample. This type of assessment can only come from a representative validation data set.

The uncertainty surfaces described in Sections 4 through 10 highlight where these biases may occur by weighting the uncertainty according to the physical and environmental proximity to points in the ASRIS database. A true assessment of performance will however only come from comparing future representative field samples to the broad predictions made here.

Historically, field observations are located according to the surveyors best judgement. Unfortunately, the precise locations or the schemes used are not always identifiable. While this has improved to some extent with the more recent use of model-based and designed-based studies (e.g. Gessler et al., 1995; McKenzie & Ryan, 1999), it needs to be adopted more pervasively if we are to be more assured of the quality of our inference in such soil property models in the future.

3.11 Model performance and future directions

The numerous sources of variation and differences in data quality made the construction of useful soil property models an inherently difficult task. Models for some properties performed well, while others performed relatively poorly. In general, topsoil models were stronger than subsoil models. There was however a large amount of unexplained variation in all models.

Property	layer	unit	N (total)	Performance on test data set				
				R ²	RMSE	abs. err	rel. err	corr.
pH	1		24319	0.67	0.77	0.56	0.51	0.82
pH	2		12193	0.54	0.96	0.72	0.59	0.74
organic carbon	1	log	11483	0.41	0.57	0.40	0.68	0.64
organic carbon	2	log	5100	0.24	0.77	0.59	0.84	0.50
total phosphorus	1	log	7377	0.62	0.92	0.68	0.54	0.79
extractable phosphorus	1	log	2124	0.35	0.85	0.67	0.78	0.59
clay content	1	sqrt	9750	0.44	1.36	1.05	0.70	0.67
clay content	2	sqrt	7050	0.22	1.60	1.21	0.86	0.47

Table 6: Model performance summary.

property	N (total)	categories	classification rate (test data)
texture layer 1	99316	5	0.535
texture layer 2	73163	3	0.671
thickness layer 1	106144	2	0.666
thickness layer 2	96384	2	0.670

Table 7: Categorical model performance summary.

Tables 6 and 7 present summaries of continuous and categorical property model performance respectively. RMSE and relative error are both unit specific and can not really be

compared across properties, though might be used to compare the performance for different layers of the same property. Note that the model performance for total nitrogen is not given here because it is not modelled directly (see Section 6).

The performance of these models varied considerably across the ASRIS extent. Some areas were however poorly represented in the database (e.g. Northern Territory, far north Queensland, western South Australia, and northern Western Australia) and, as such, have predictions which must necessarily be treated with more caution.

Some residual spatial structure often existed at distances less than 5km. This suggests that while the models may capture the longer range variation, the short range variation is not always adequately explained. An examination of more local models may help improve this, though preliminary investigations in the Burdekin and Katanning areas found difficulties in obtaining adequately predictive models. There is definitely some scope to make a more direct account for the spatial dependence in this type of modelling.

The relative importance of the environmental variables to the models and the ensuing predicted surfaces is not discussed in this document. A detailed analysis of the rules derived is still in progress at the time of writing this report. This is an involved task because individual variables may feature in different parts of the decision tree. Moreover, interpreting the meaning of a rule is complicated by the fact that combinations of environmental variables may actually be used as surrogates for other effects. The climatic variates do however feature prominently in all models. The most used terrain attributes are generally elevation, relative elevation and relief.

Other future areas of work might be to investigate if other environmental variables, such as gamma radiometric remote sensing or additional terrain variables, improve the predictive ability. Effort could also be put towards a further cleaning of the database, with a view to increasing the homogeneity by reducing the sources of variation. Using more data is not necessarily better, and reducing the data set to one of a higher overall standard may actually result in soil property models giving improved predictions and greater applicability.

4 pH

Laboratory assessed soil pH is one of the most well-populated soil property fields in the database. As measured pH values were known to be dependent on the assessment method and the soil to solution ratio, only methods 4A1 (pH in water, ratio 1:5), 4B1 and 4B2 (pH in CaCl_2 , ratio 1:5) were considered in the point modelling. For a full description of these methods see Rayment & Higginson (1992).

Measurements in water were converted to equivalent recordings in CaCl_2 according to the calibration equation detailed in the appendix.

Preparation of the layer 1 and 2 pH data involved making unit corrections to pH values recorded as $\text{pH} \times 10$, deleting some numeric missing value flags, and omitting some observations.

4.1 Layer 1 pH in CaCl_2

There were 25915 layer 1 pH measurements available. Tables 8 and 9 break these down over the States/CSIRO and methods. The column labels min and max refer to the minimum and the maximum respectively, while q10, q25, q50, q75 and q90 denote the 10th, 25th, 50th, 75th and 90th quantiles in turn.

State	min	q10	q25	q50	q75	q90	max	N
NSW	2.8	4.0	4.3	4.8	5.5	6.4	8.7	2591
QLD	3.3	4.6	5.1	5.8	6.5	7.3	8.6	4642
SA	3.7	5.5	6.5	7.7	8.4	8.8	10.2	3930
TAS	3.4	4.0	4.4	4.9	5.4	5.9	8.2	656
VIC	3.5	4.4	4.9	6.5	7.7	7.9	8.7	1308
WA	3.2	4.1	4.4	4.7	5.1	5.7	8.5	11808
CSIRO	3.2	4.1	4.5	4.9	5.9	6.8	8.0	980

Table 8: Distribution by State/CSIRO.

Method	min	q10	q25	q50	q75	q90	max	N
4A1	2.8	4.3	4.8	5.5	6.6	7.6	8.8	10848
4B1	2.8	4.0	4.4	4.7	5.0	5.7	9.3	11978
4B2	3.2	5.3	6.5	7.8	8.5	8.8	10.2	3089

Table 9: Distribution by pH method.

The histograms of layer 1 pH by State are given in Figure 4. The most obvious features are the large number of counts for Western Australia, their relative acidity, and the relative alkalinity of South Australia.

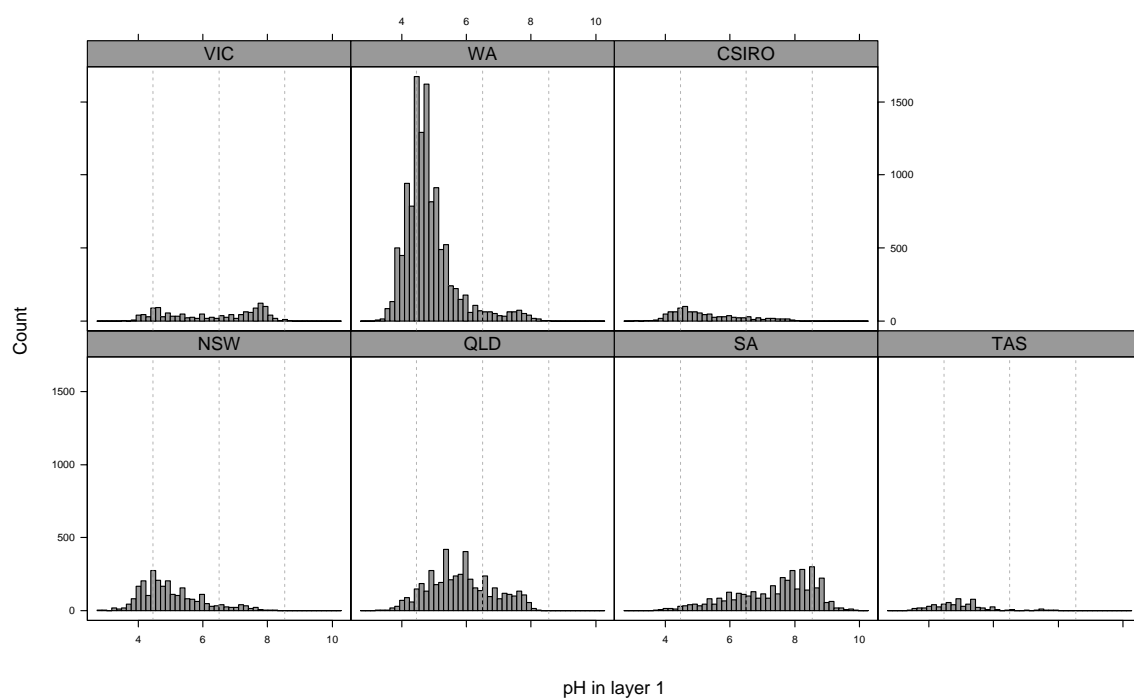


Figure 4: Histograms of layer 1 pH in CaCl_2 by State.

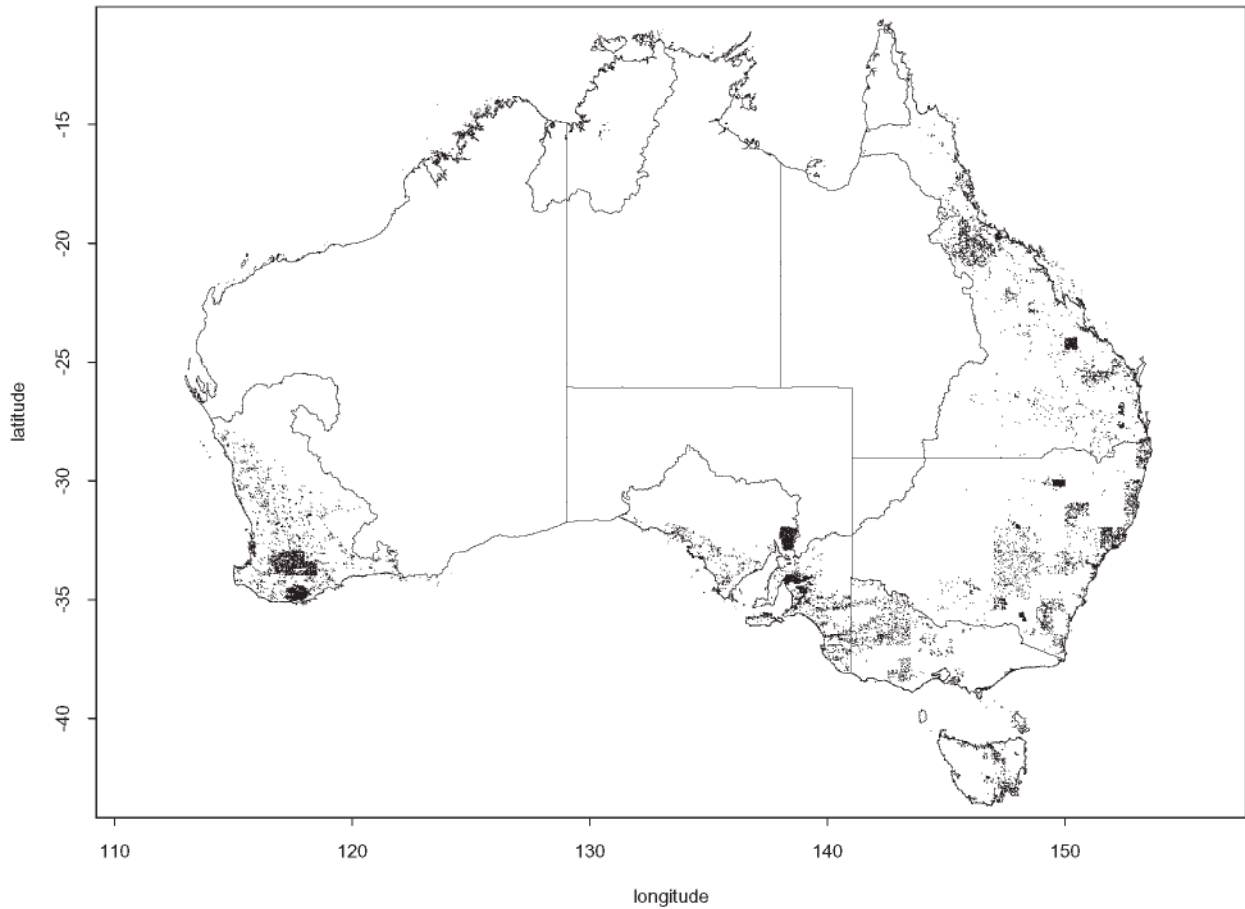


Figure 5: Locations of layer 1 pH observations.

The locations of the layer 1 pH observations used in the modelling are given in Figure 5. There is a good dispersion of points across the states, though clearly a large degree of variation in the sampling intensity.

A *Cubist* piecewise-linear model was fitted to these data. 30 variables were used: 11 climatic, 3 MSS, 14 terrain, lithology and landuse. 24319 points were available inside the ASRIS extent and with all environmental predictors. The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.67	0.77	0.56	0.51	0.82

Table 10: pH in layer 1 model diagnostics on test data set.

The overall performance of this model on the test data is summarized in Figure 6. While there remains considerable unexplained variation, the model clearly demonstrates valuable predictive ability.

The most notable feature in the residual plots is that there are many pH values that the model substantially over-estimates. This is largely attributable to South Australia where the model does poorly at predicting values with lower pH (See Figure 8). The quantile plot and histogram of the residuals for the test data are given in Figure 7.

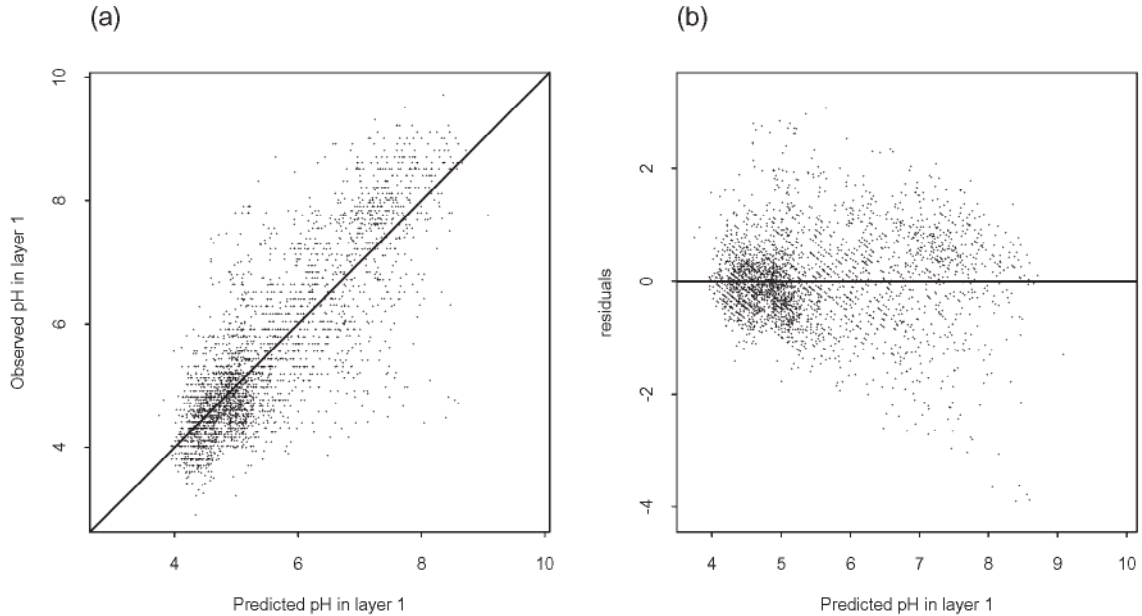


Figure 6: (a) Observed versus predicted and (b) residual plot for pH in layer 1 model (test data only). Note only a random sample of 3000 points are plotted.

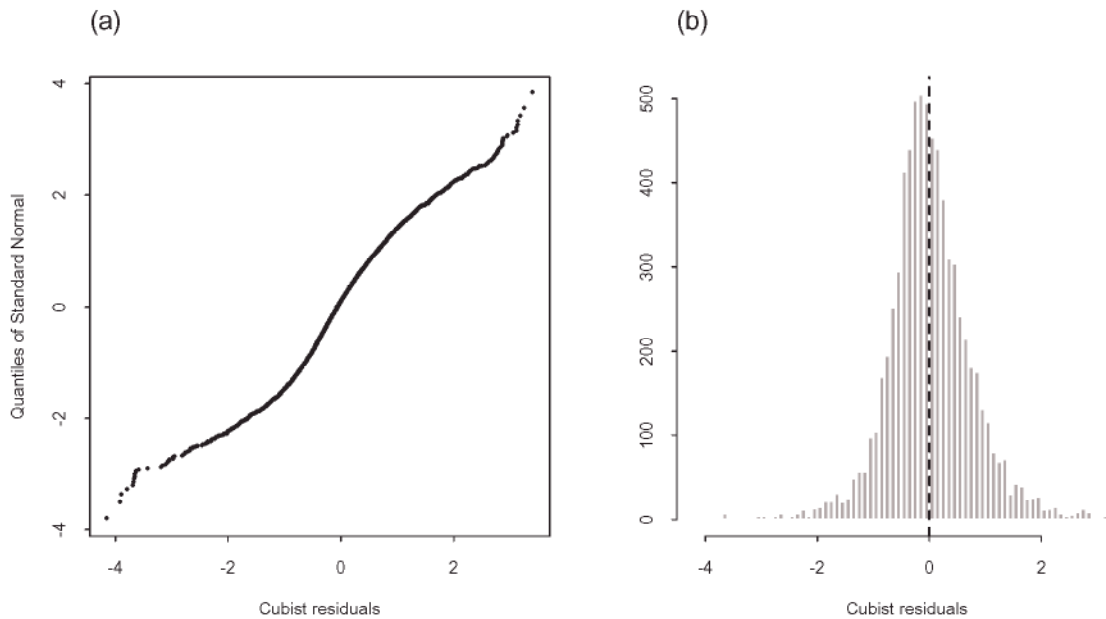


Figure 7: (a) Quantile plot and (b) histogram of Cubist residuals for pH in layer 1 model (test data).

The model was then refitted using the same model options and variables on all 24319 observations. 27 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 0.56; relative error 0.51; correlation 0.82).

The relative performance of the fitted model for all the layer 1 pH data across the individual states and regions can be assessed in Tables 11 and 12. Note that all measures are calculated within the State or region with the predicted values given by the Australia-wide model.

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
NSW	2465	0.63	0.72	0.55	0.72
QLD	4329	0.66	0.70	0.57	0.73
SA	3830	0.45	0.87	0.90	1.12
TAS	522	0.51	0.85	0.45	0.63
VIC	1283	0.81	0.46	0.59	0.75
WA	10935	0.55	0.85	0.43	0.62
CSIRO	955	0.74	0.62	0.52	0.70

Table 11: Performance of pH in layer 1 model by state/CSIRO.

Region	N	rank correlation	relative error	average error	approximate RMSE
nterritory	41	-0.21	1.21	0.83	0.97
carpentaria	13	0.39	0.91	0.65	0.79
qldnorth	63	0.62	0.57	0.27	0.35
qldcentral	819	0.76	0.65	0.43	0.58
qldsouth	542	0.49	0.89	0.54	0.69
moreton	274	0.25	1.03	0.52	0.68
burdekin	1364	0.47	0.84	0.60	0.76
fitzroy	704	0.47	0.85	0.63	0.81
mdb	4296	0.81	0.53	0.65	0.83
nswnorth	481	0.38	0.96	0.44	0.58
nswsouth	745	0.44	0.85	0.43	0.62
viceast	148	0.37	0.94	0.52	0.70
vicwest	747	0.76	0.54	0.77	0.98
tasmania	521	0.51	0.85	0.45	0.63
sagulf	2410	0.44	0.87	0.93	1.16
wplateau	226	0.33	0.99	0.85	1.07
wasouth	10562	0.55	0.86	0.43	0.61
indian	363	0.21	0.85	0.56	0.75

Table 12: Performance of pH in layer 1 model by Region.

Figure 6 can be decomposed into the 18 regions that make up the ASRIS extent. This leads directly to Figures 8 and 9. A maximum of 2500 points are plotted per panel for

clarity. If there are more in any ASRIS region, a random sample of 2500 points was taken. Some care should be taken in interpreting these figures as the relative point density is not always evident and can be misleading at this resolution. For example, it is difficult to appreciate the greater density of points with high pH in the *sagulf* panel.

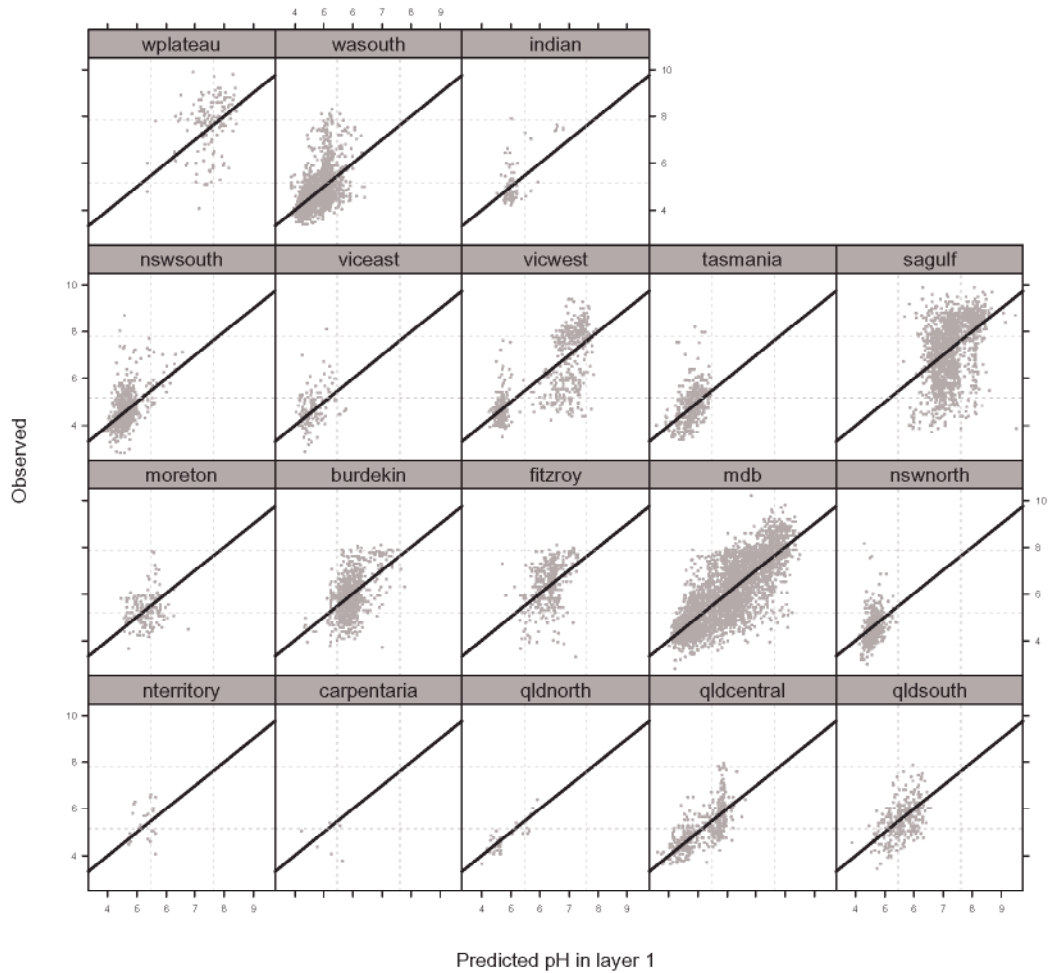


Figure 8: Observed versus predicted plots by region.

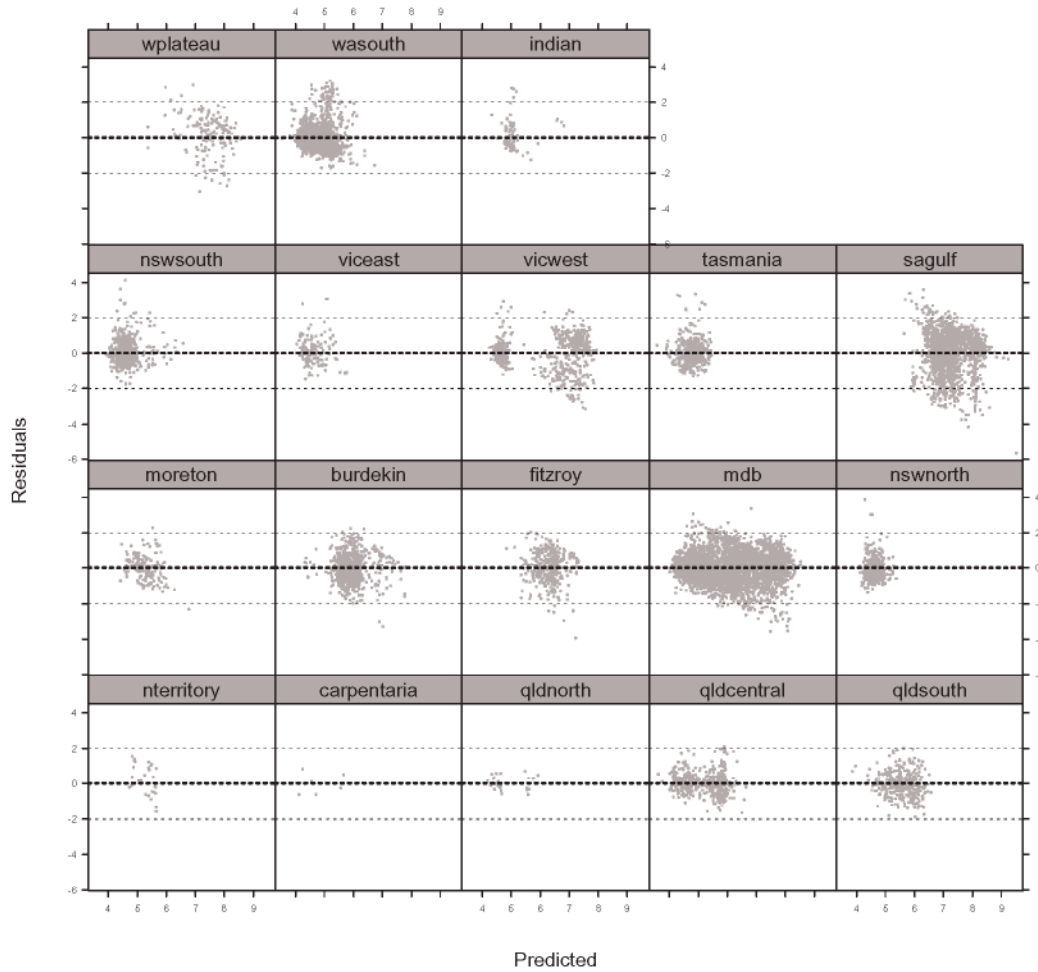


Figure 9: Residual plots by region.

The diagnostic plots and summary statistics suggest that the model generally performs well in the Murray-Darling basin, south/central Queensland. The overall performance in Western Australia and Tasmania is fair. Coastal New South Wales and the Moreton region are predicted fairly poorly. The performance in South Australia is particularly poor with a tendency to over-estimate the low pH values recorded.

The model residuals were examined for spatial structure. The omni-directional State-based variograms are given in Figure 10. The latitudes/longitudes used are those of the 250m pixel centroids. There is clearly residual spatial structure. The ranges are however short with little structure left outside a distance of 0.05 degrees (approximately 5km). This suggests that the Cubist model has not managed to explain all the short range variation but has captured the longer range signal. Some explanation for this can be found in the fact that the climatic variables proved to be good predictors but only have a resolution of 5km. The large variation in the sill should be noted, e.g. compare South Australia and Western Australia. This reflects the general poorer performance in South Australia and subsequently greater variation in model residuals.

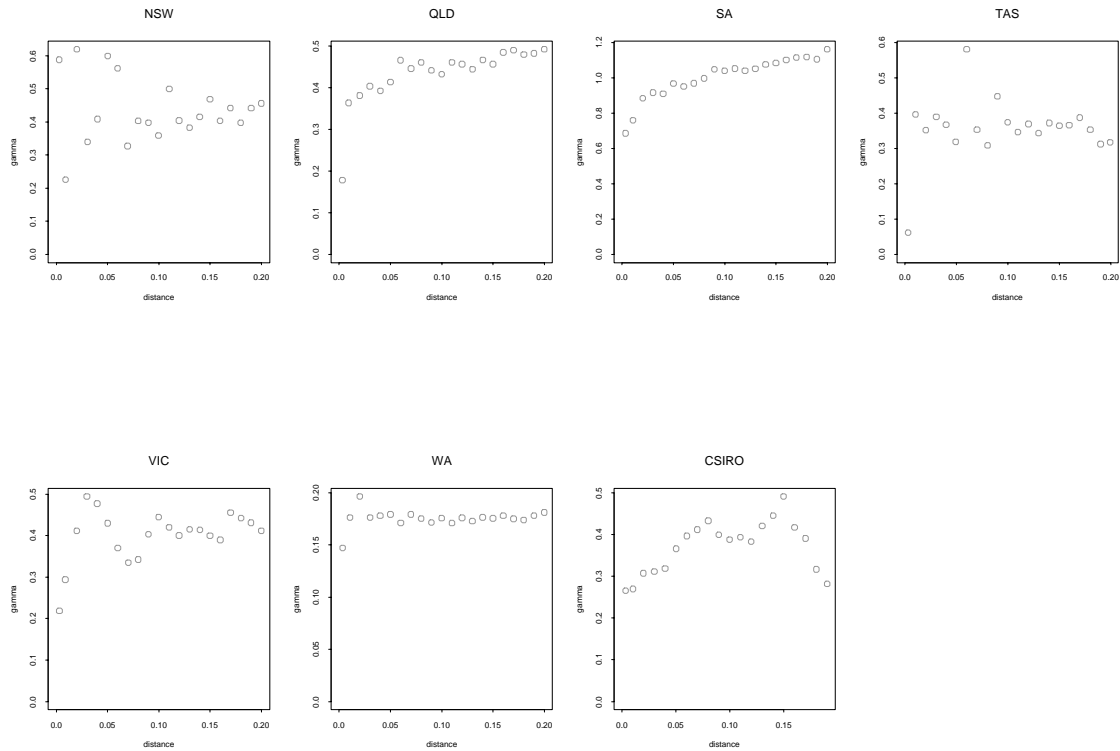


Figure 10: Residual variograms by State/CSIRO (distance in degrees).

The rules derived from the final Cubist model were applied to the ASRIS extent to generate a map of predicted layer 1 pH. This surface is illustrated in Figure 11. The digital maps for all soil property predictions are available at www.nlwra.gov.au/data/.

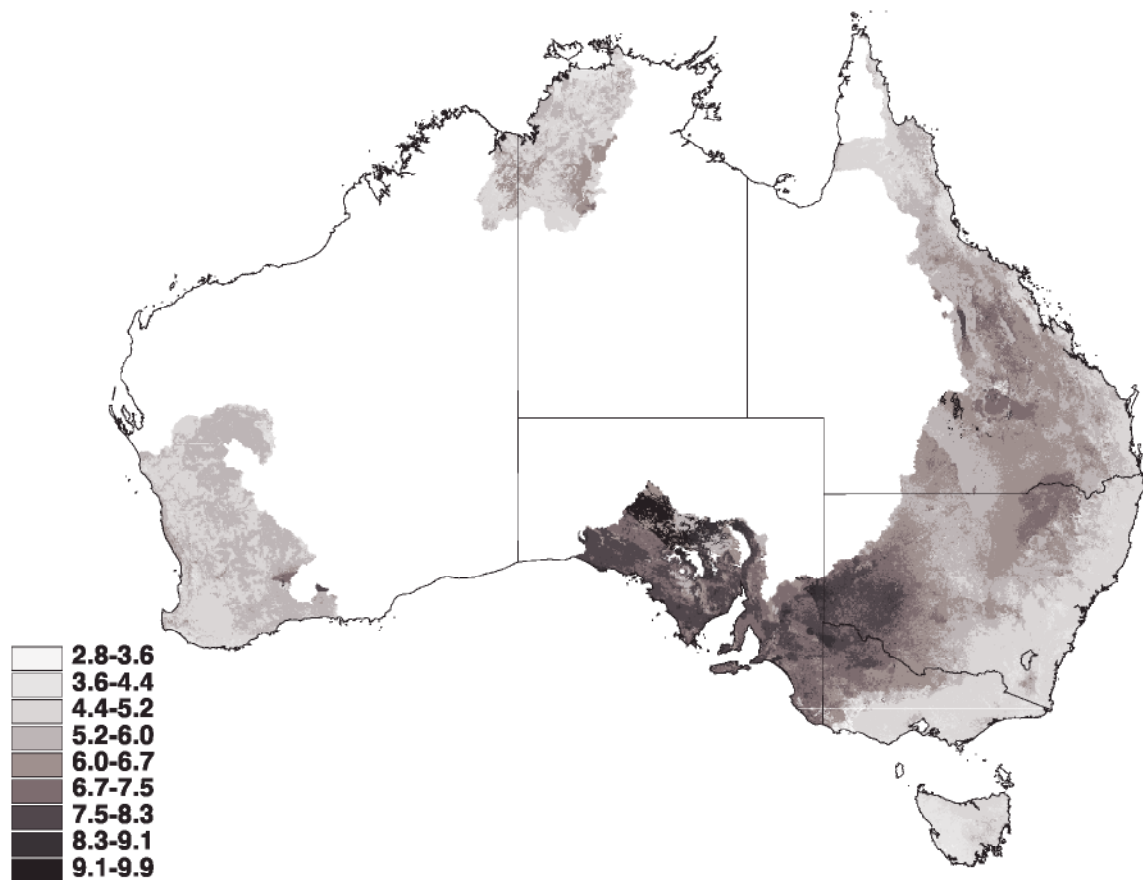


Figure 11: pH prediction surface

A model certainty surface was made for this prediction surface according to the method described in Section 3.8. Figures 12, 13 and 14 give the three component surfaces, namely the log point density, the environmental representativeness and the regional model performance. Larger values for all surfaces are associated with greater model certainty.

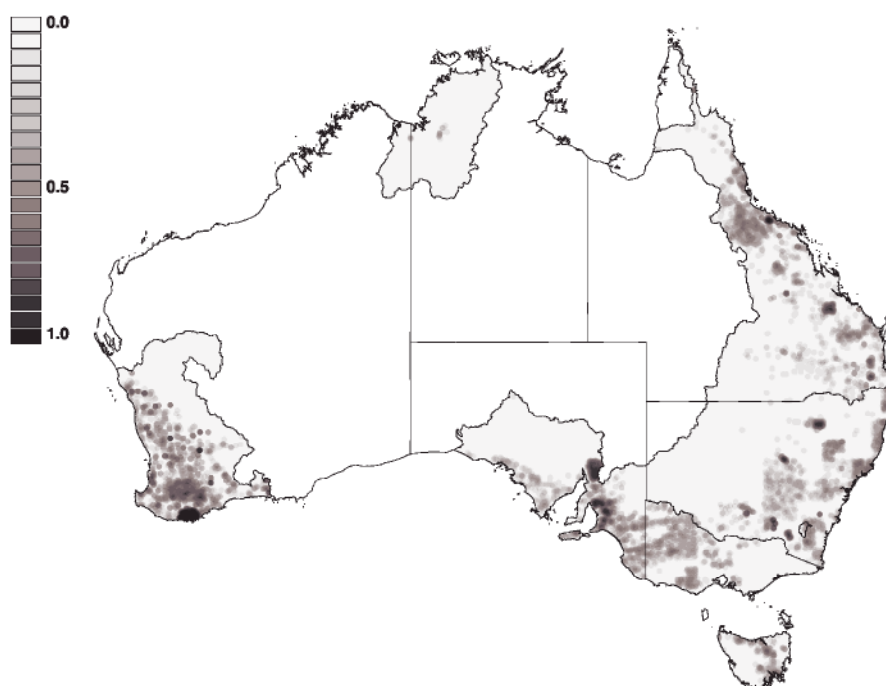


Figure 12: Smoothed log point density

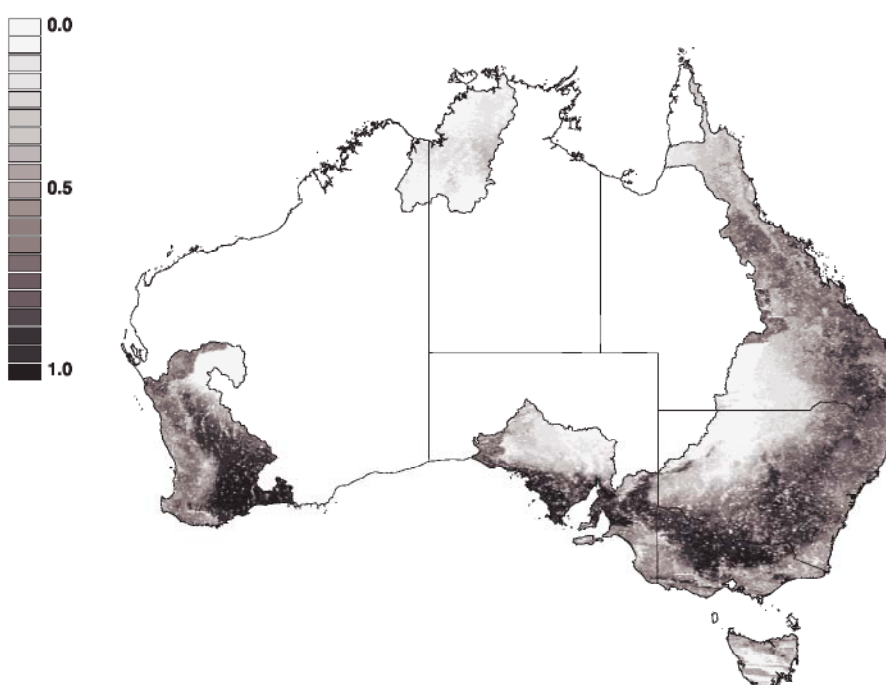


Figure 13: Environmental representativeness

8 variables were used to make the environmental representativeness surface, namely: elevation, relief, relative elevation, MSS band 2, annual mean moisture index, maximum temperature of the warmest period, precipitation of the warmest period and the highest period radiation. It is evident in Figure 13 that there is greater extrapolation into the Northern Territory, north-eastern Western Australia, northern South Australia, north-western New South Wales and south-western Queensland.

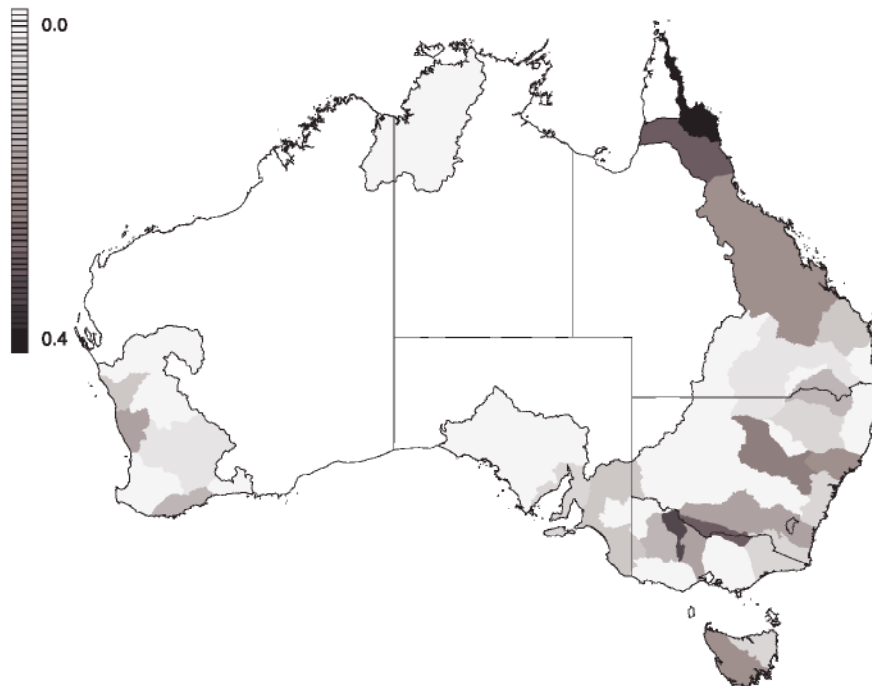


Figure 14: 1 – relative error

A weighted average of these three component surfaces was taken and gives the overall certainty surface. This surface is illustrated in Figure 15. While the polygonal structure shows through in the certainty surface, boundaries are less severe than in Figure 14 because they are now modified by the log point density and the environmental representativeness components. This certainty surface, and those of other models, can also be viewed at www.nlwra.gov.au/data/.

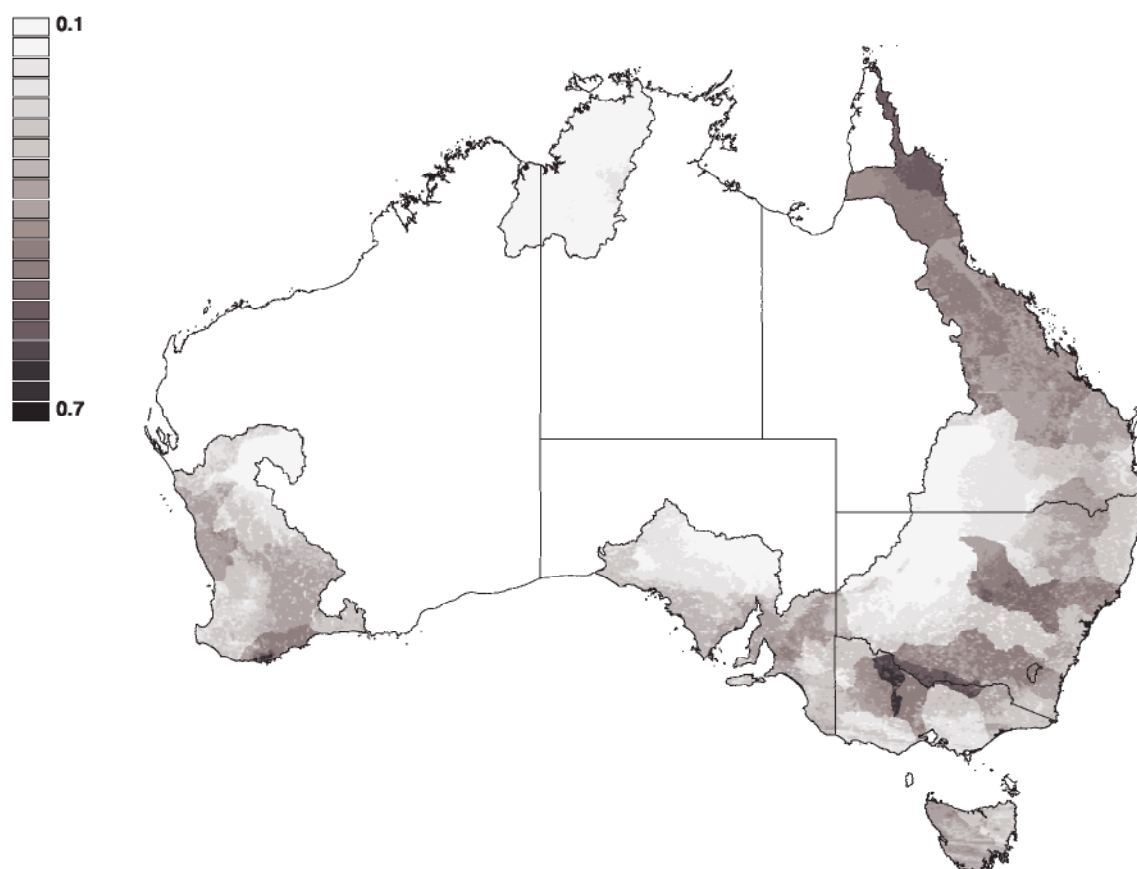


Figure 15: Layer 1 pH certainty surface

4.2 Incorporating data from Theme 5 Project 4D

NLWRA Project 5.4D examines the “Nutrient balance in regional farming systems and soil nutrient status” and provides an alternative source of pH measurements. These data were investigated in South Australia due the poor performance of the continental pH model there.

Figure 16 presents the histogram of the layer 1 pH in CaCl_2 values according to the 4 ASRIS regions intersecting in South Australia. There are clearly a small number of low pH values in the *wplateau* region. The *sagulf* region appears bi-modal with a clusters of both low and high measured pH. The *mdb* and *vicwest* regions are much more evenly distributed.

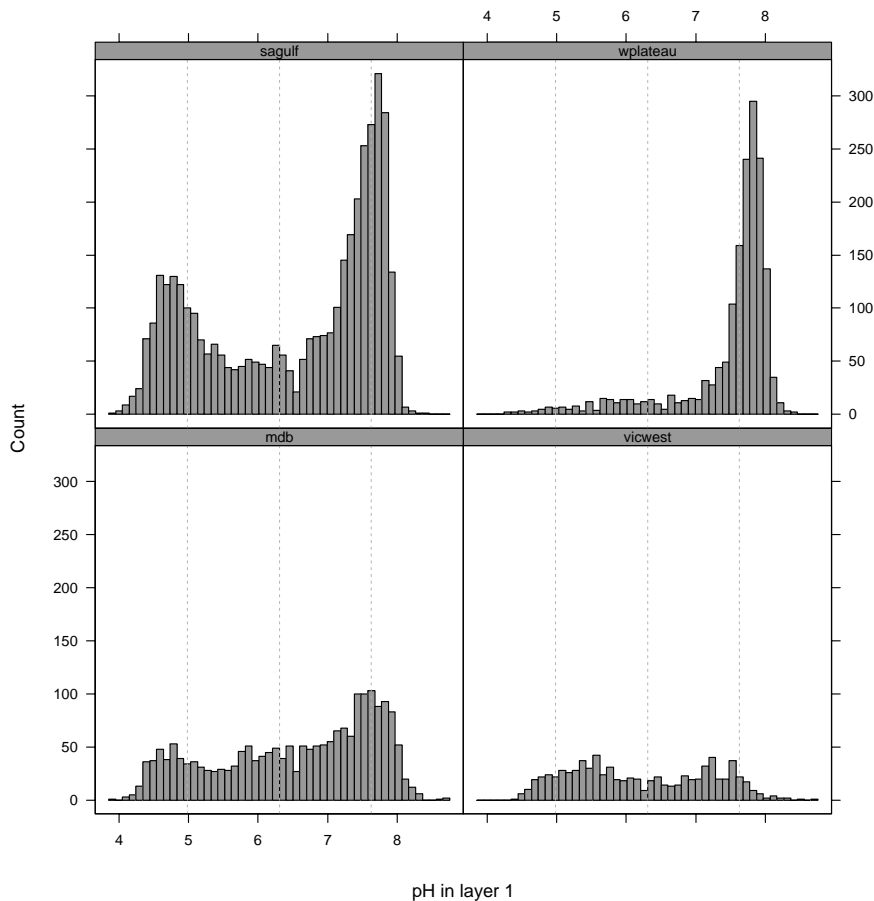


Figure 16: Histograms of Theme 5.4D layer 1 pH in CaCl_2 by ASRIS region in South Australia.

These data are however imperfectly located as they are only geo-referenced to the nearest “100” in order to preserve anonymity. To apply the methodology thus far described requires us to disaggregate these pH values, from their geo-referenced “100s” to their surrounding districts, so as to associate environmental predictors at a more local level.

This was done as follows:

- Thiessen polygons were created around the coverage of the 8959 unique “100” locations.
- landuse and texture coverages were overlaid to divide each Thiessen polygon into a number of smaller zones. Each zone was identified by specific landuse and texture combinations. A total of 29032 zones was created.
- the pH values, tagged to each “100” (Thiessen polygon), were then disaggregated to the centroids of these smaller zones according to their known texture and landuse classes. For example, if an observation had a landuse code 12 and a texture class 5 then its location was disaggregated to the centroid of the zone with landuse code 12 and a texture class 5. Where there was more than one zone with the same landuse and texture class combination within a Thiessen polygon, the pH value(s) were allocated randomly with probabilities determined by the relative zonal area. Observations with landuse classes and/or texture classes not used in the texture and landuse coverages were not disaggregated.
- A suite of environmental variables was then associated with each disaggregated point observation. These were obtained by calculating the zonal mean (or mode if categorical) for each environmental variable over the zone(s) with that specific landuse/texture combination in the appropriate Thiessen polygon.
- A disaggregated data set of the form

[ID, disaggregated location, pH value, (average) environmental variables ...]

was then created. 19422 disaggregated points were available, however only 16282 of these points had layer 1 pH measurements reported.

A *Cubist* piecewise-linear was fitted to the disaggregated data. 40 variables were used: 19 climatic, 4 MSS, 15 terrain, lithology (1:2m) and landuse. There were 16282 points inside the extent with all environmental predictors. The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R ²	RMSE	average error	relative error	correlation
0.52	0.82	0.62	0.59	0.72

Table 13: pH in layer 1 model diagnostics on Theme 5.4D test data set.

The overall performance of this model on the test data is summarized Figure 17. It is evident that the predictive power is not strong as there is a large degree of variability in the predictions made. There is some intensive clustering at the high pH values and to a lesser extent at low measured pH. Despite the clustering, there is a clear tendency to over-predict some low pH values and under-predict some high pH values. The quantile plot and histogram of the residuals for these test data are given in Figure 18.

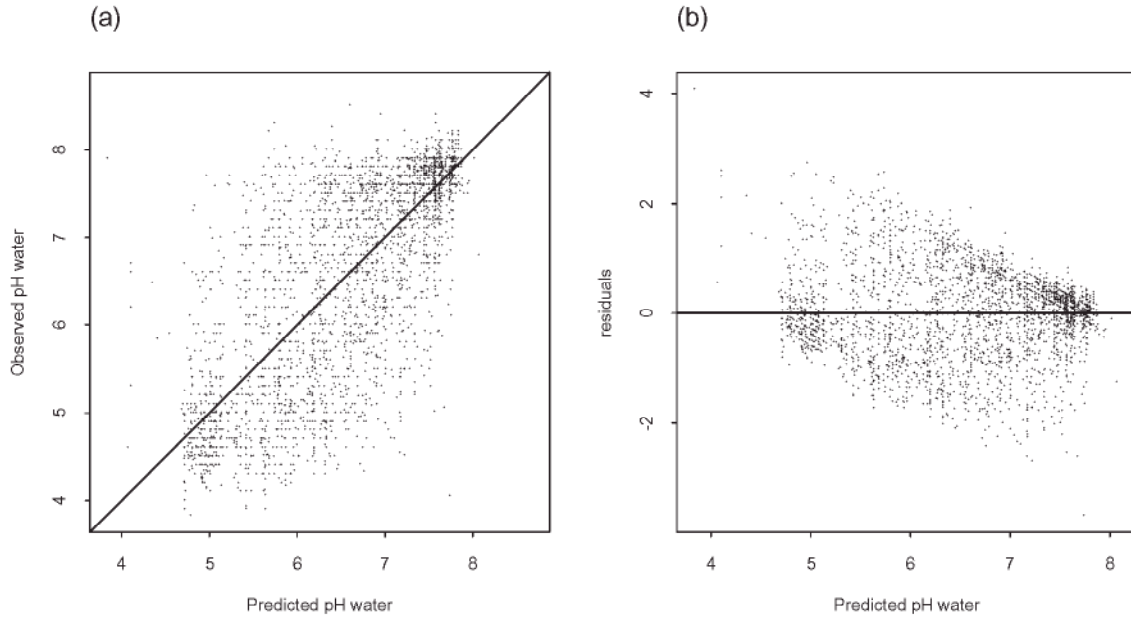


Figure 17: (a) Observed versus predicted and (b) residual plot for pH in layer 1 model (test data only). Note only a random sample of 3000 points are plotted.

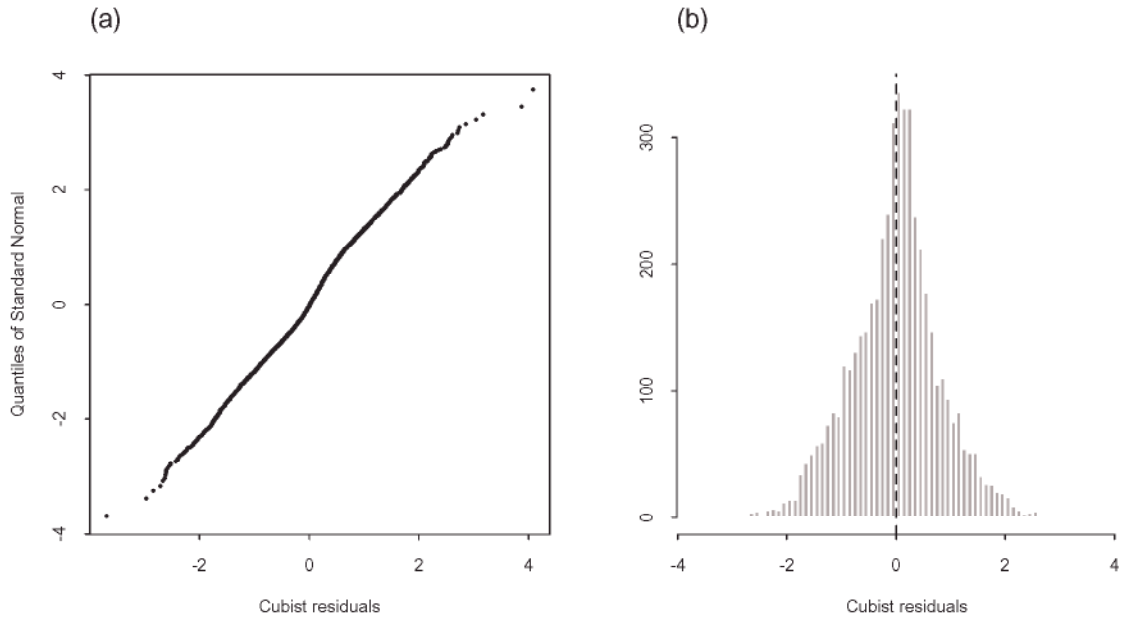


Figure 18: (a) Quantile plot and (b) histogram of Cubist residuals for pH in layer 1 model. Test data only.

The model was then refitted using the same model options and variables on all 16282 observations. 23 rules were used in the model. 10-fold cross-validation was performed

on this model to judge performance (average error 0.61; relative error 0.57; correlation 0.74). The relative performance of the fitted model by ASRIS region can be assessed in Table 14.

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
mdb	3240	0.68	0.63	0.60	0.75
vicwest	3574	0.47	0.85	0.82	0.98
sagulf	7076	0.75	0.50	0.53	0.72
wplateau	2392	0.67	0.64	0.37	0.56

Table 14: Performance of pH in layer 1 model by region.

Figure 17 can be decomposed into the 4 regions that make up South Australian ASRIS extent. This leads directly to Figures 19 and 20.

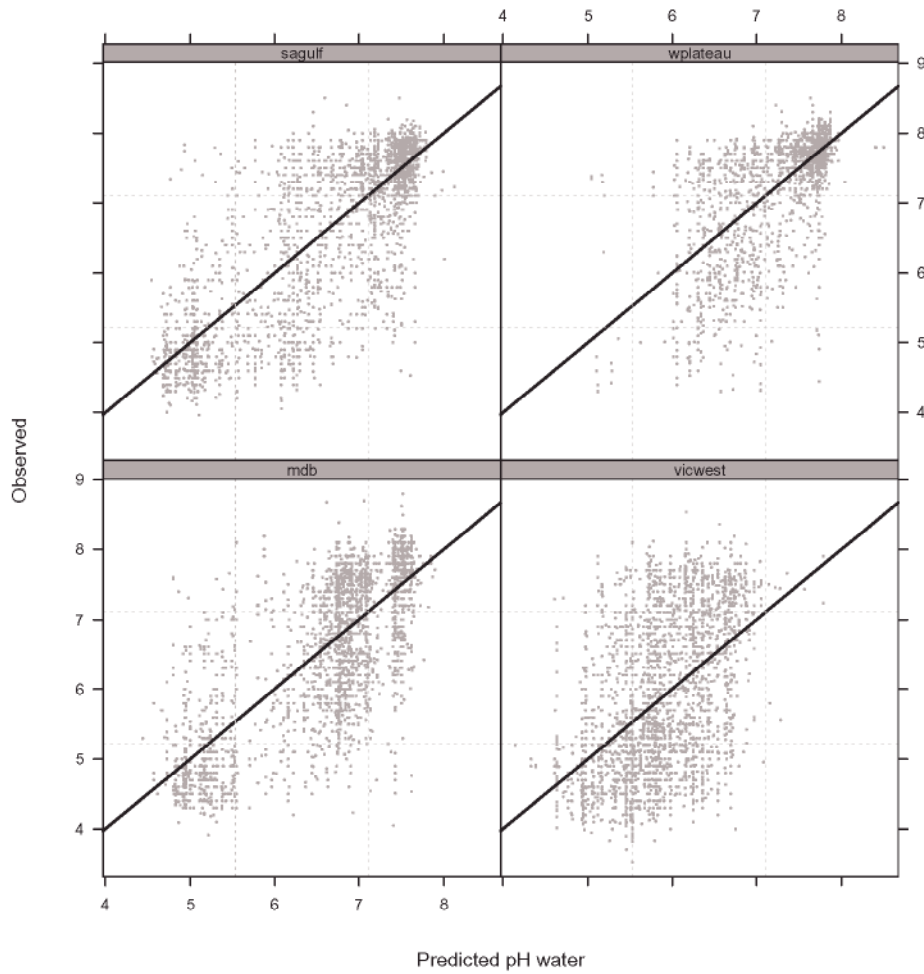


Figure 19: Observed versus predicted plots by region.

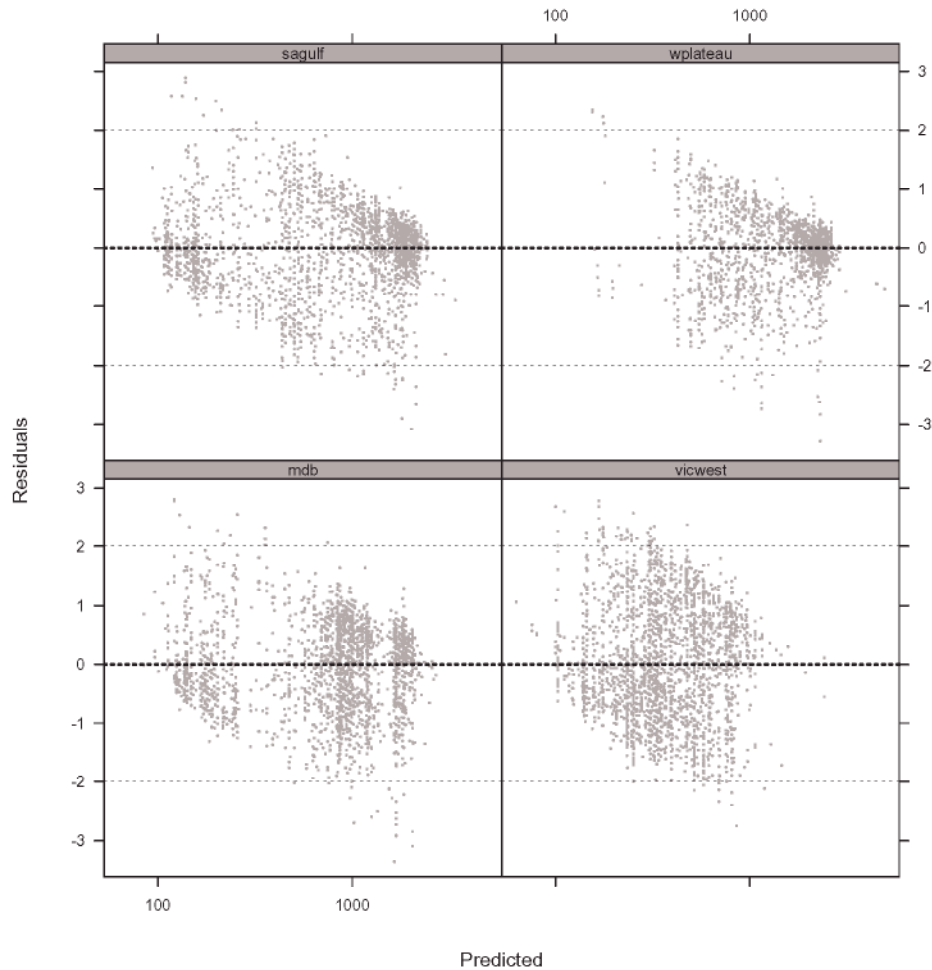


Figure 20: Residual plots by region.

It is evident from the decomposition in Figures 19 and 20 and the diagnostics in Table 14 that the model performance is better in the *sagulf* and *mbd* regions, and that *vicwest* demonstrates poor predictive ability. Some structure is present in the residual plots in Figure 20.

The rules derived from this model were applied to that part of South Australia within the ASRIS extent to generate an alternative map of layer 1 pH. This map can be seen at www.nlwra.gov.au/data.

Comparing the model predictions from this data to the ASRIS predictions identifies some differences. The predictions from this model are more acidic on Kangaroo Island and the Fleurieu peninsula. The predictions in the north of the State and on the Eyre peninsula are however more acidic than expected. Both of these regions do however represent areas where there are fewer points and thus a greater degree of extrapolation.

4.3 Layer 2 pH in CaCl_2

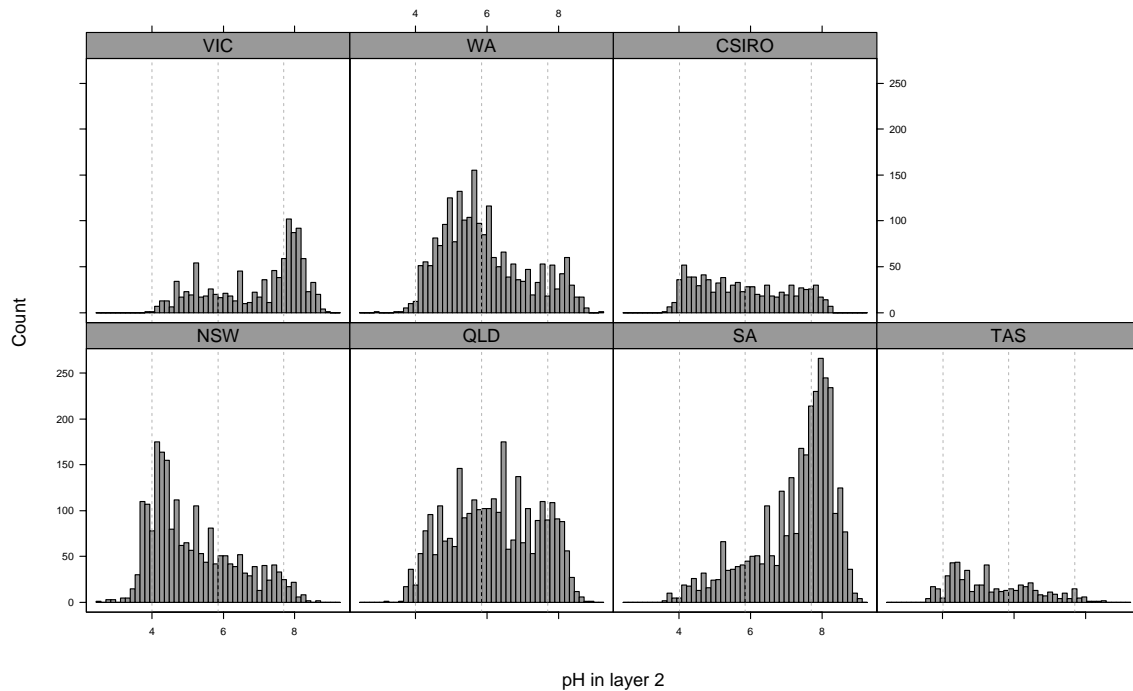
There were 12810 layer 2 pH measurements available. Tables 15 and 16 break these down over the State/CSIRO and methods.

State	min	q10	q25	q50	q75	q90	max	N
NSW	2.5	3.9	4.2	4.8	5.9	7.1	8.7	2120
QLD	3.2	4.5	5.3	6.2	7.2	7.9	8.9	2957
SA	3.6	5.5	6.7	7.7	8.1	8.4	9.1	3068
TAS	3.7	4.1	4.5	5.2	6.2	7.1	8.5	541
VIC	3.9	5.0	5.8	7.5	8.0	8.3	8.9	1053
WA	2.9	4.5	5.0	5.7	6.6	7.9	9.2	2187
CSIRO	3.6	4.2	4.7	5.6	6.9	7.7	8.2	884

Table 15: Distribution by State/CSIRO.

Method	min	q10	q25	q50	q75	q90	max	N
4A1	2.5	4.4	5.2	6.5	7.7	8.2	9.1	10160
4B1	2.8	4.2	4.8	5.5	6.4	7.6	9.2	2454
4B2	3.8	4.0	4.1	4.3	4.8	5.2	7.4	196

Table 16: Distribution by pH method.

Figure 21: Histograms of layer 2 pH in CaCl_2 by State.

The histograms of pH in layer 2 in Figure 21 demonstrate the relative alkalinity of South Australia and Victoria and the relative acidity of New South Wales, Western Australia and Tasmania. The locations of the layer 2 pH observations used in the modelling are given in Figure 22.

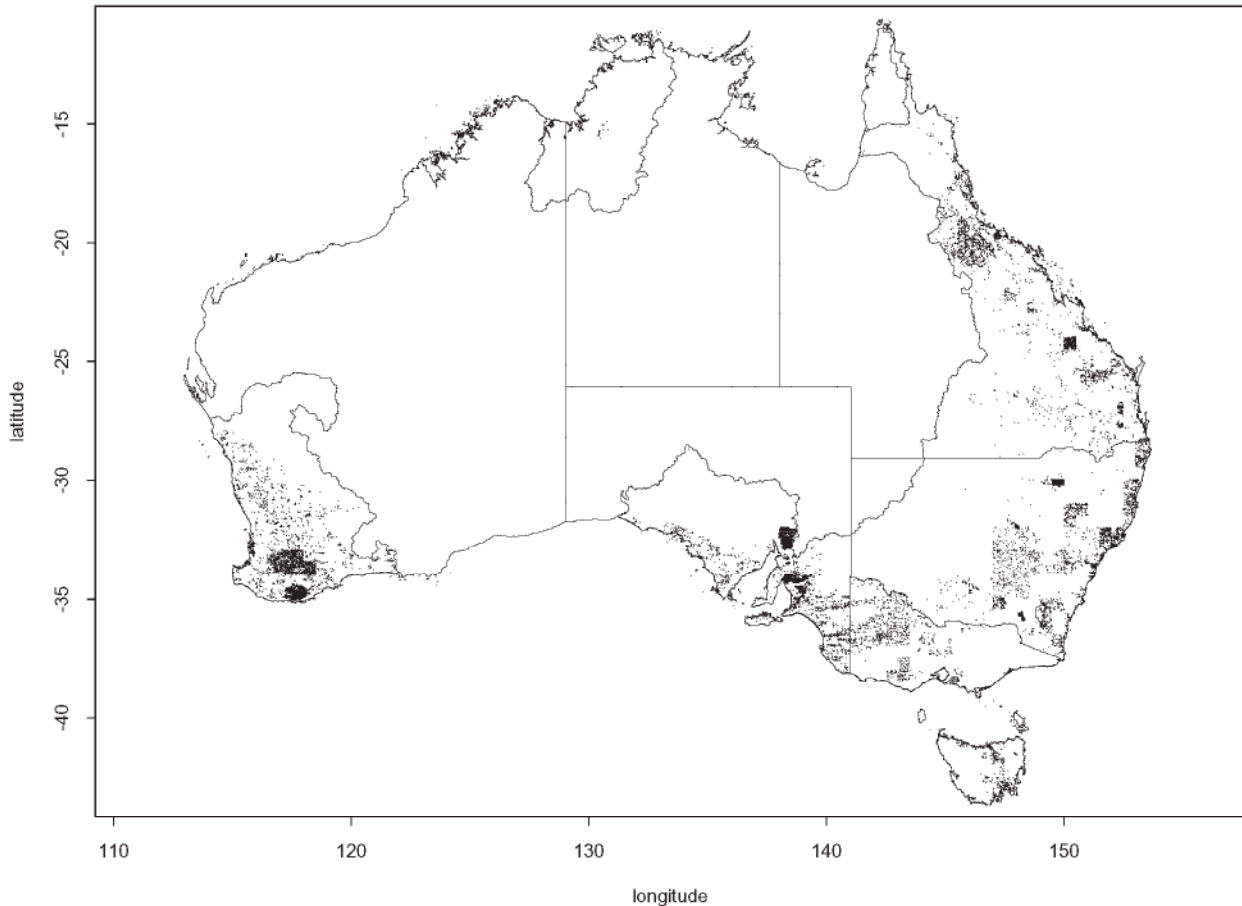


Figure 22: Locations of layer 2 pH observations.

A *Cubist* piecewise-linear was fitted to these data. 30 variables were used: 11 climatic, 3 MSS, 14 terrain, lithology and landuse. 12193 points were available inside the extent with all environmental predictors. The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.54	0.96	0.72	0.59	0.74

Table 17: pH in layer 2 model diagnostics on test data set.

The overall performance of this model on the test data is summarized graphically in Figure 23. While there is again a large degree of scatter and thus considerable unex-

plained variability, the model obviously exhibits reasonable predictive power. It is however notably weaker than the layer 1 pH model. All the diagnostics in Table 17 are less favourable.

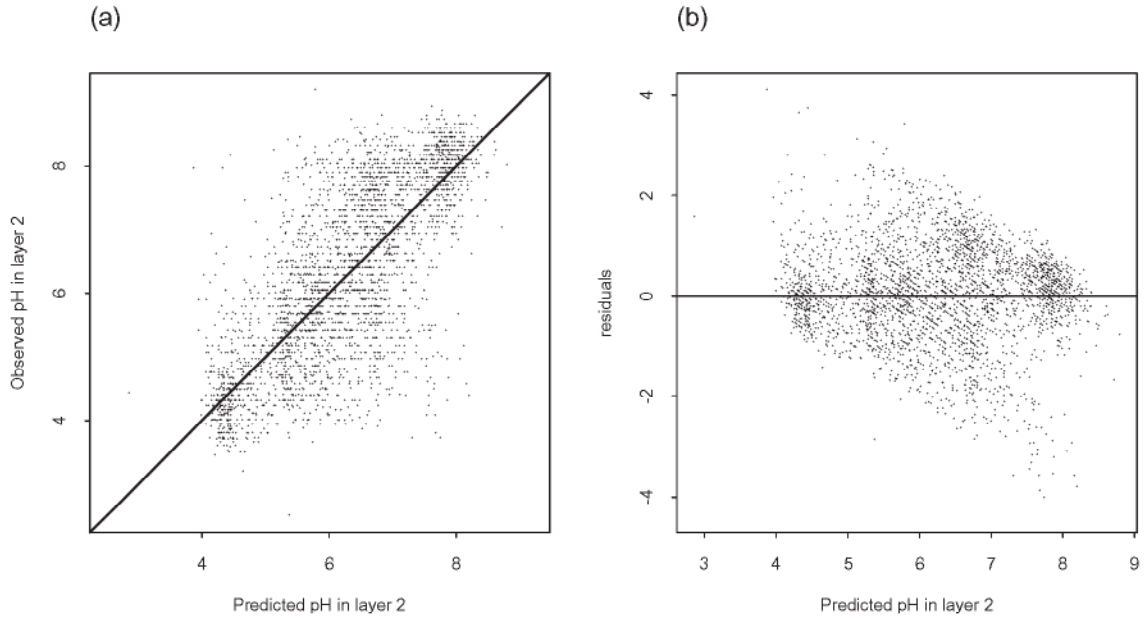


Figure 23: (a) Observed versus predicted and (b) residual plot for pH in layer 2 model (test data only). Note only a random sample of 3000 points are plotted.

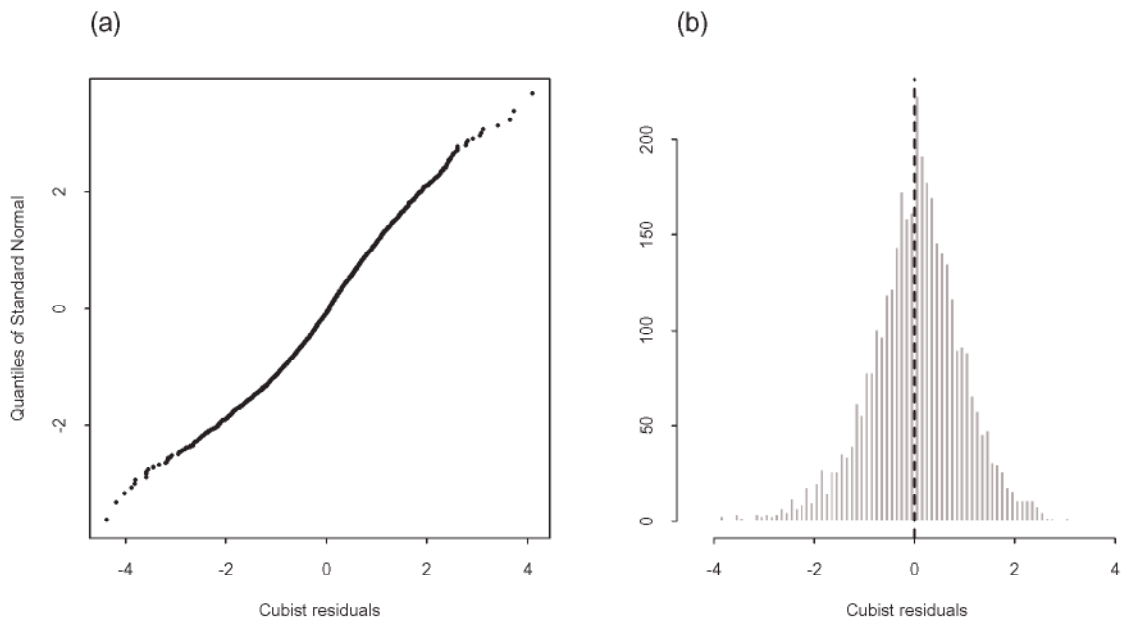


Figure 24: (a) Quantile plot and (b) histogram of Cubist residuals for pH in layer 2 model (test data).

The most notable feature in the residual plots is those low pH values that the model over-estimates. Similarly to layer 1 pH, this can be largely be attributed to the poor performance in South Australia. The quantile plot and histogram of the residuals for the test data are given in Figure 24.

The model was then refitted using the same model options and variables on all 12193 observations. 27 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 0.68; relative error 0.55; correlation 0.77).

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
NSW	2025	0.70	0.60	0.60	0.80
QLD	2781	0.68	0.66	0.68	0.87
SA	3024	0.59	0.77	0.69	0.94
TAS	434	0.62	0.72	0.64	0.83
VIC	1035	0.81	0.43	0.49	0.69
WA	2031	0.45	0.85	0.82	1.04
CSIRO	863	0.80	0.52	0.59	0.77

Table 18: Performance of pH in layer 2 model by state/CSIRO.

Region	N	rank correlation	relative error	average error	approximate RMSE
nterritory	37	0.53	0.76	0.77	0.92
carpentaria	9	0.74	1.19	0.64	0.82
qldnorth	26	0.62	0.61	0.40	0.54
qldcentral	610	0.82	0.55	0.52	0.69
qldsouth	223	0.54	0.82	0.78	0.98
moreton	160	0.24	0.92	0.78	0.97
burdekin	1121	0.42	0.89	0.73	0.91
fitzroy	303	0.23	0.98	0.77	0.94
mdb	3213	0.81	0.47	0.59	0.79
nswnorth	438	0.25	0.94	0.47	0.67
nswsouth	699	0.49	0.78	0.54	0.75
viceast	110	0.47	0.84	0.52	0.66
vicwest	654	0.61	0.68	0.75	0.97
tasmania	434	0.62	0.72	0.64	0.83
sagulf	2083	0.63	0.75	0.69	0.95
wplateau	52	0.21	1.42	0.73	0.91
wasouth	1984	0.45	0.85	0.81	1.03
indian	37	0.19	1.13	1.03	1.27

Table 19: Performance of pH in layer 2 model by region.

The relative performance of the fitted model for all the layer 2 pH data across the individual States and regions can be assessed in Tables 18 and 19. Note that all measures are calculated within the State or region with the predicted values given by the Australia-wide model.

Figure 23 can be decomposed into the 18 regions that make up the ASRIS extent. This leads directly to Figures 25 and 26.

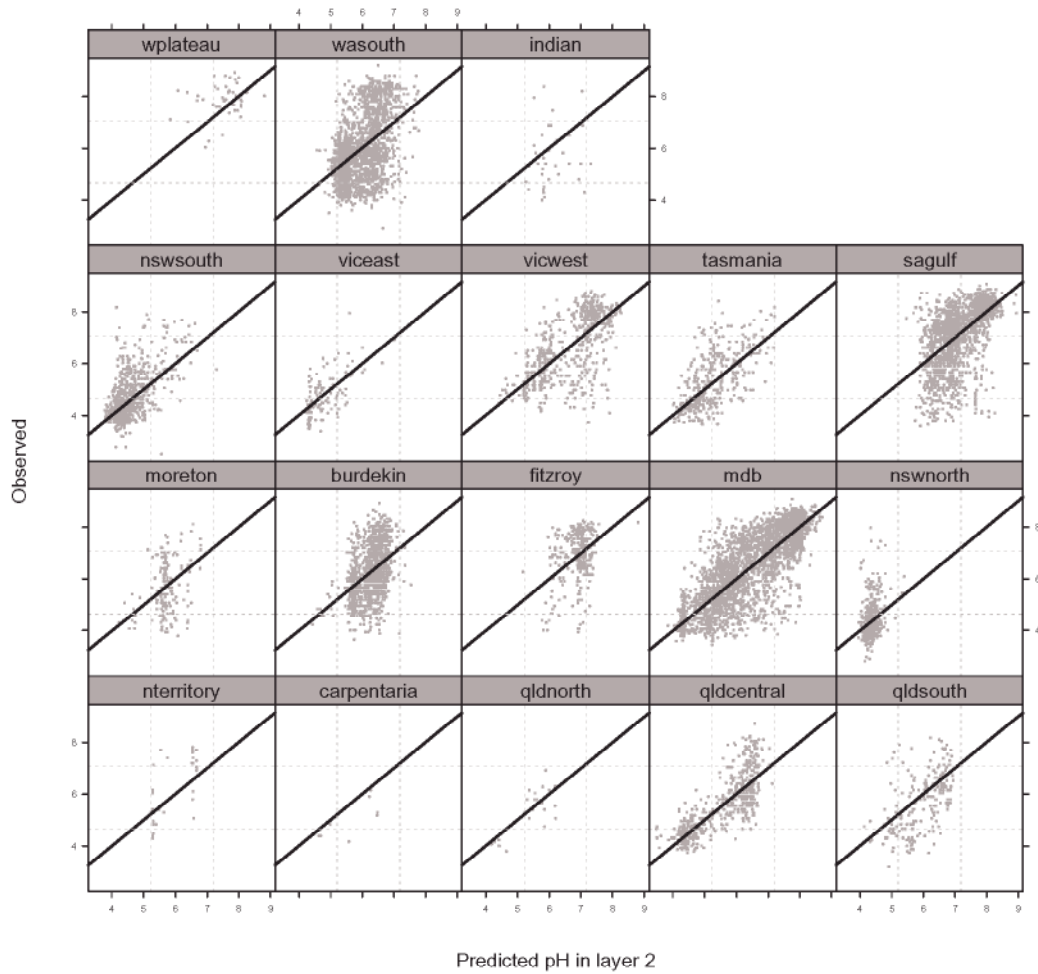


Figure 25: Observed versus predicted plots by region.

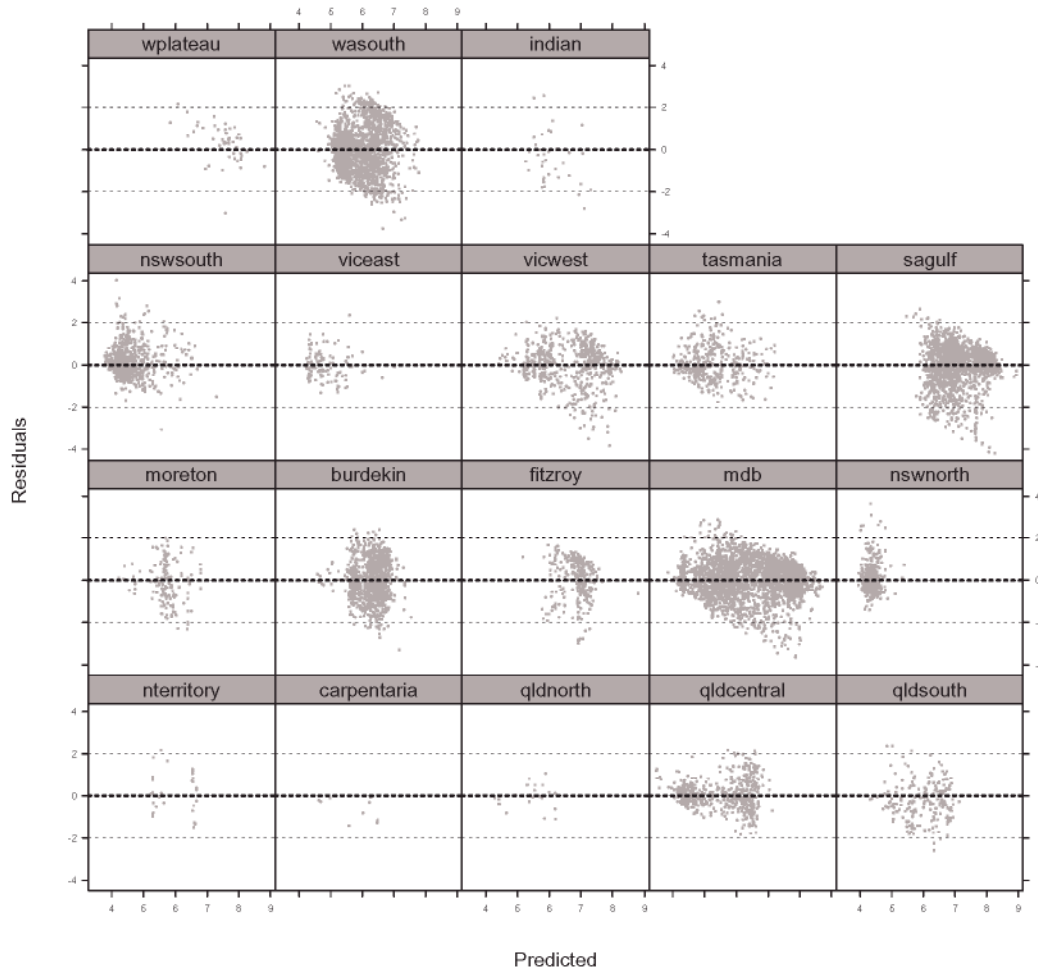


Figure 26: Residual plots by region.

The diagnostic plots and summary statistics suggest that the model generally performs well in the Murray-Darling basin, south/central Queensland and Tasmania. The overall performance in southern New South Wales, Victoria, Fitzroy and Burdekin is fair. Northern New South Wales and South Australia are predicted fairly poorly.

The residual variograms in Figure 27 again exhibit some unaccounted spatial dependence, particularly in South Australia and Victoria.

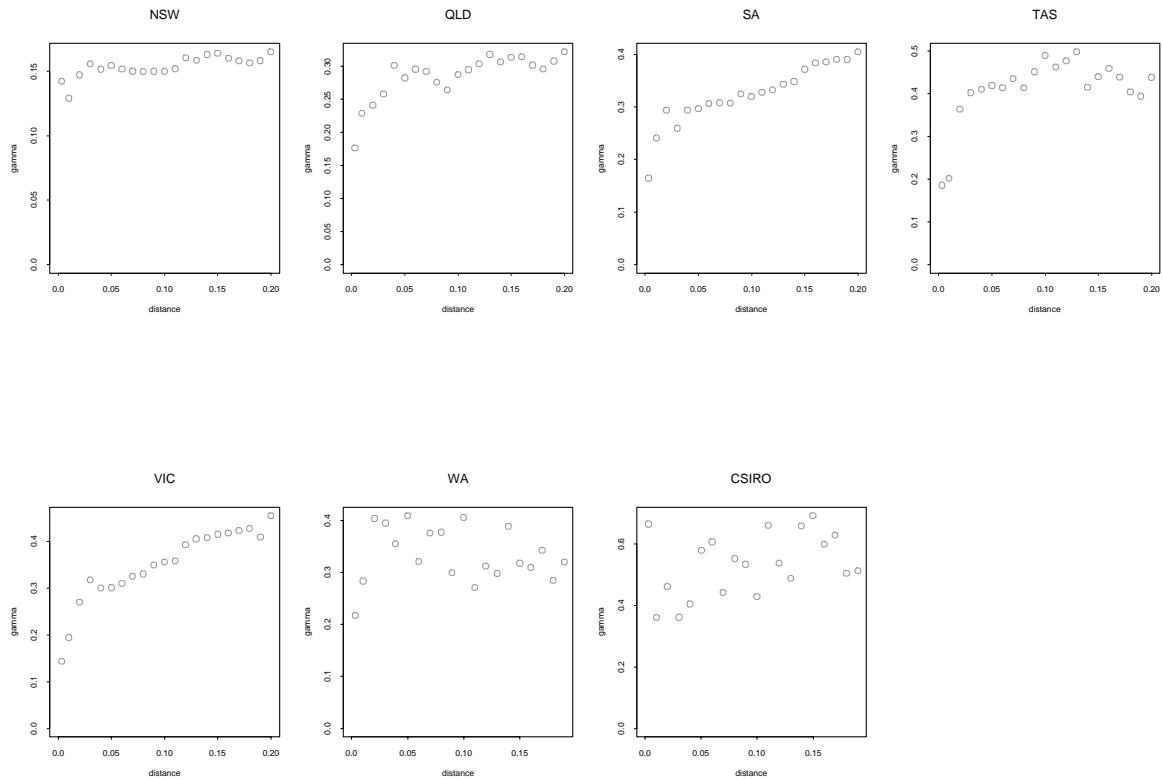


Figure 27: Residual variograms by State/CSIRO (distance in degrees).

The 27 rules from the final Cubist model were applied to ASRIS extent to generate a map of layer 2 pH predictions. This map can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. 8 variables were used to make the environmental representativeness surface, namely: elevation, relief, relative elevation, MSS band 2, annual mean moisture index, maximum temperature of the warmest period, precipitation of the warmest period and the highest period radiation. This certainty surface can be found at www.nlwra.gov.au/data.

5 ORGANIC CARBON

The organic carbon content is widely used to assess the amount of organic matter in soil. Rayment & Higginson (1992) remark that while the levels of organic carbon are typically higher in surface soils, the amount of organic matter is highly variable, though usually less than 5%.

There are two broad methods for assessment of organic carbon. The Walkley and Black method (Walkley, 1947) and that attributed to Heanes (1984) use wet oxidation, while the second involves the combustion of the sample in a high frequency induction furnace.

The Walkley-Black method is generally known to give incomplete recovery, with the recovery often historically quoted in the vicinity of 75-80% (Rayment & Higginson, 1992). While a correction factor from the incomplete Walkley and Black methods to total organic carbon of 1.3 (i.e. $\approx 1/0.8$) is sometimes offered as a rule of thumb, there is no universal correction factor.

In an Australia-wide investigation Skjemstad et al. (2000) found that differences existed that could be attributed to the state (laboratory) and the date at which the sample was analysed, with more recent analyses showing a much more complete recovery. The appropriate correction factor was notably less than 1.3 and for a large part of the data not needed at all. In light of this, and not actually being able to determine whether the correction factor had already been applied or not to the ASRIS data, it was decided that no correction factor would be used.

Those methods considered in the ASRIS point modelling were the Walkley and Black (6A1, 6A1.UC), Heanes wet oxidation (6B1) and the combustion methods (6B2, 6B3 and 6.DC). A detailed description of each of these procedures can be found in Rayment & Higginson (1992). All methods were assumed to estimate total organic carbon.

5.1 Layer 1 organic carbon

The data as summarized by State/CSIRO and method are presented in Tables 20 and 21. Histograms of organic carbon by State are given in Figure 28. It is clear that the distributions are strongly right-skewed. Prior to modelling some large organic carbon values were omitted. These are not displayed in Tables 20 and 21. A natural log transformation was found to be useful in reducing some of the skewness and stabilizing the variance.

State	min	q10	q25	q50	q75	q90	max	N
NSW	0.01	0.73	1.19	2.10	3.49	5.41	14.59	2438
QLD	0.02	0.63	0.90	1.40	2.20	3.30	14.00	3595
SA	0.05	0.50	0.80	1.30	2.00	3.03	13.70	738
TAS	0.04	1.51	2.70	4.38	6.40	9.30	14.70	461
VIC	0.10	0.60	0.90	1.40	2.40	4.29	14.02	634
WA	0.04	0.89	1.61	2.78	3.39	3.86	14.80	3378
CSIRO	0.20	1.00	1.37	2.17	4.60	7.87	14.49	828

Table 20: Distribution by State/CSIRO.

Method	min	q10	q25	q50	q75	q90	max	N
6A1	0.02	0.64	0.92	1.40	2.20	3.40	14.02	3934
6A1.UC	0.04	0.85	1.54	2.76	3.45	4.37	14.80	3913
6B1	0.01	0.75	1.28	2.25	3.63	5.52	14.59	2157
6B2	0.10	0.60	1.10	1.80	5.07	8.09	14.49	585
6B3	0.05	0.80	1.09	1.63	2.50	3.98	13.70	1132
6.DC	0.04	0.68	1.00	2.60	4.93	6.94	14.70	351

Table 21: Distribution by organic carbon method.

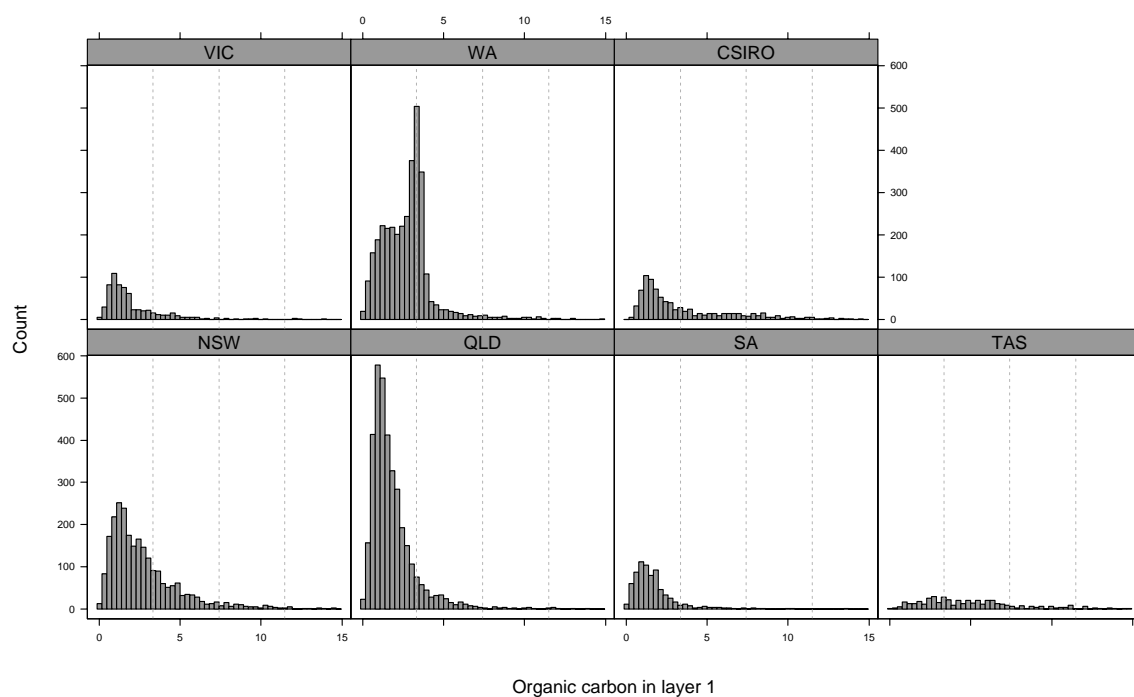


Figure 28: Histograms of layer 1 % organic carbon by State.

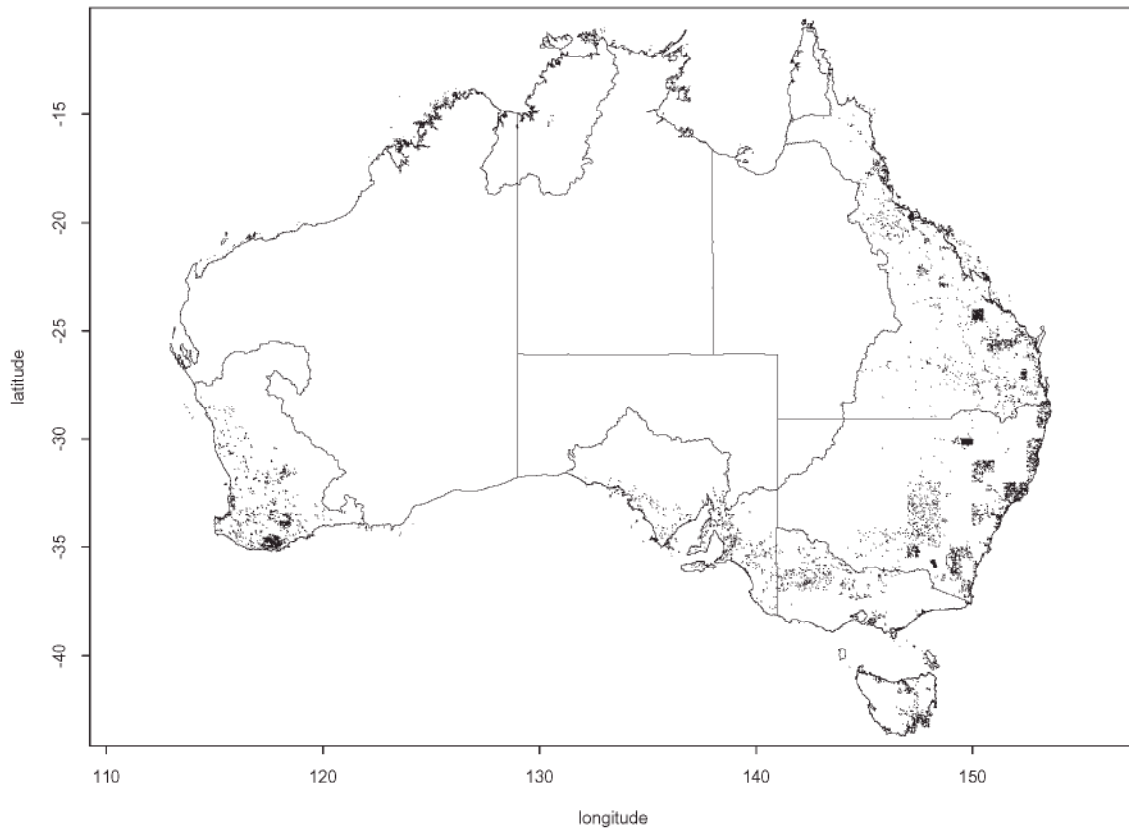


Figure 29: Locations of layer 1 % organic carbon observations.

The locations of the 11483 organic carbon observations used in the modelling are given in Figure 29.

A Cubist piecewise linear was fitted to the natural logarithm of these data. 30 variables were used: 11 climatic, 3 MSS, 13 terrain, lithology, landuse and ASC. The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.41	0.57	0.40	0.68	0.64

Table 22: $\log(\text{organic carbon})$ in layer 1 model diagnostics on test data set.

The overall performance on the test data is summarized in Figure 30. The residuals appear well distributed. Looking at the observed versus predicted figure there is evidently some tendency to over-estimate low $\log(\text{\% organic carbon})$ and under-estimate the higher $\log(\text{\% organic carbon})$.

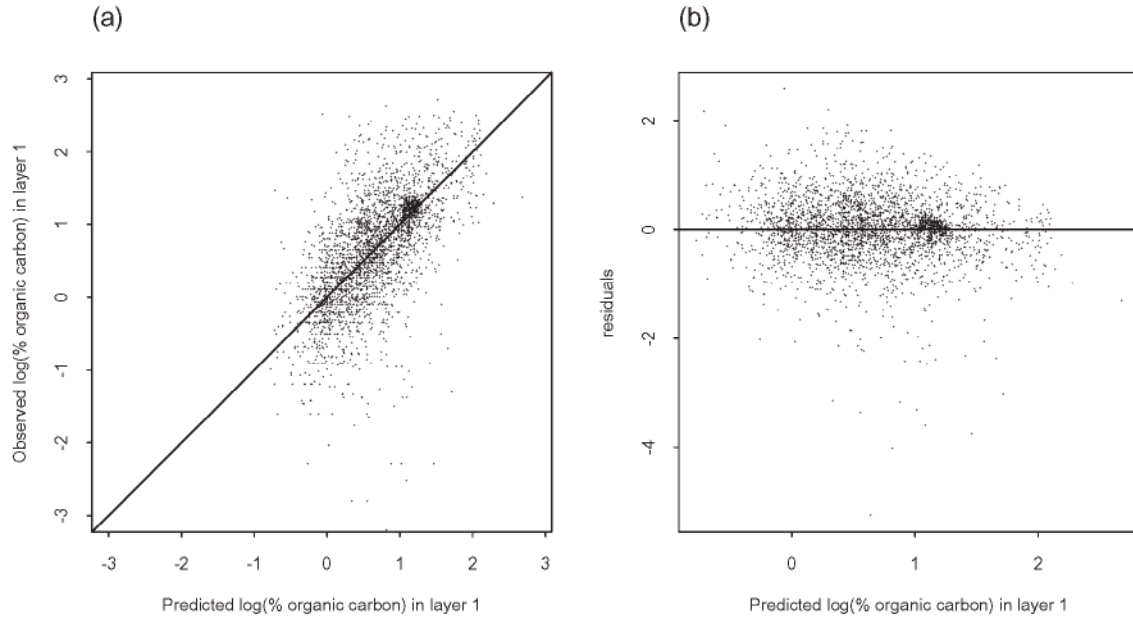


Figure 30: (a) Observed versus predicted and (b) residual plot for $\log(\% \text{ organic carbon})$ in layer 1 model (test data only). Note only a random sample of 3000 points is plotted.

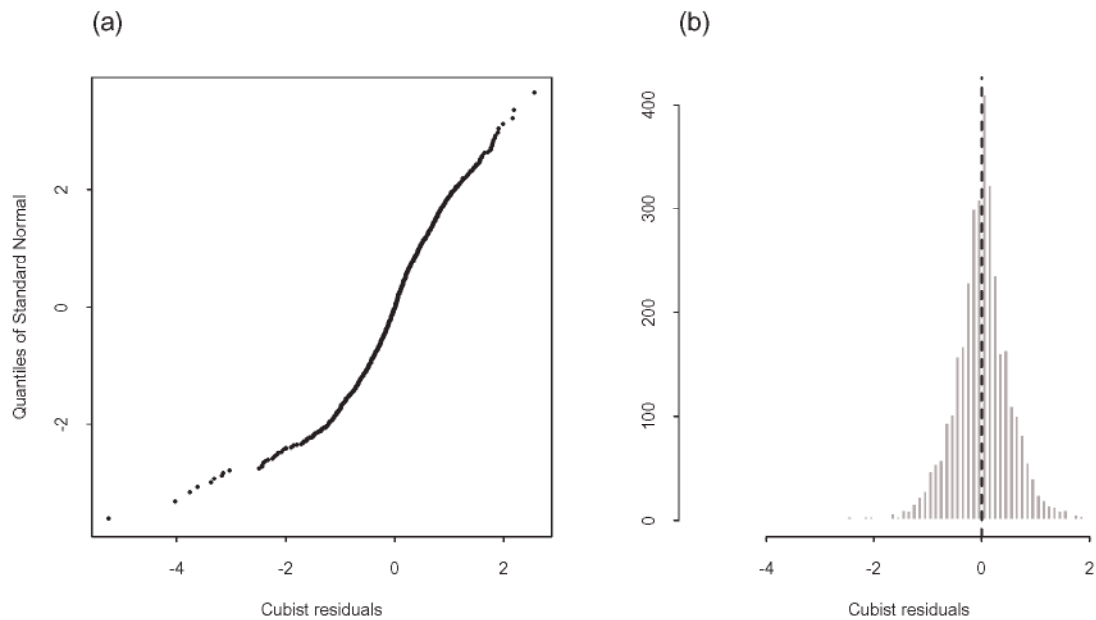


Figure 31: (a) Quantile plot and (b) histogram of Cubist residuals for $\log(\% \text{ organic carbon})$ in layer 1 model (test data).

The quantile plot and histogram of residuals are given in Figure 31. The presence of some very small assessments of % organic carbon values explain the negative skewness that can be evidenced in both Figures 30 and 31.

The model was then refitted using the same model options and variables on all 11483 observations. 29 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 0.39; relative error 0.65; correlation 0.68).

The state and region-wise performance are summarized in the Tables 23 and 24.

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
NSW	2340	0.58	0.78	0.50	0.69
QLD	3343	0.65	0.73	0.37	0.49
SA	706	0.63	0.78	0.45	0.62
TAS	382	0.58	0.79	0.45	0.61
VIC	615	0.72	0.66	0.39	0.55
WA	3290	0.72	0.58	0.30	0.46
CSIRO	807	0.78	0.55	0.37	0.49

Table 23: Performance of layer 1 organic carbon model by state/CSIRO.

Region	N	rank correlation	relative error	average error	approximate RMSE
nterritory	20	0.36	0.93	0.47	0.77
carpentaria	10	0.62	0.47	0.32	0.42
qldnorth	62	0.61	0.75	0.41	0.66
qldcentral	741	0.74	0.61	0.35	0.47
qldsouth	588	0.43	0.87	0.40	0.53
moreton	224	0.29	0.95	0.43	0.57
burdekin	456	0.52	0.83	0.35	0.45
fitzroy	656	0.58	0.81	0.31	0.40
mdb	3033	0.67	0.68	0.41	0.55
nswnorth	473	0.37	0.93	0.60	0.84
nswsouth	764	0.41	0.91	0.52	0.71
viceast	229	0.50	0.85	0.47	0.63
vicwest	135	0.32	1.02	0.51	0.75
tasmania	381	0.57	0.79	0.45	0.62
sagulf	364	0.63	0.79	0.41	0.55
wplateau	67	0.62	0.87	0.44	0.58
wasouth	3245	0.71	0.60	0.30	0.46
indian	35	0.37	1.05	0.44	0.61

Table 24: Performance of layer 1 organic carbon model by region.

Figure 30 can be decomposed into the 18 regions making up the ASRIS extent. This leads directly to Figure 32 and Figure 33 and enables some spatial assessment of performance.

The performance appears strongest in central/northern Queensland, eastern South Australia, the Murray-Darling basin and southern Western Australia. In coastal New South

Wales, Victoria and southern Queensland however the performance is not as strong. Some regions are not well enough represented in terms of point data for their performance to be considered too deeply, e.g. the Northern Territory, far north Queensland, Carpentaria, and northern Western Australia.

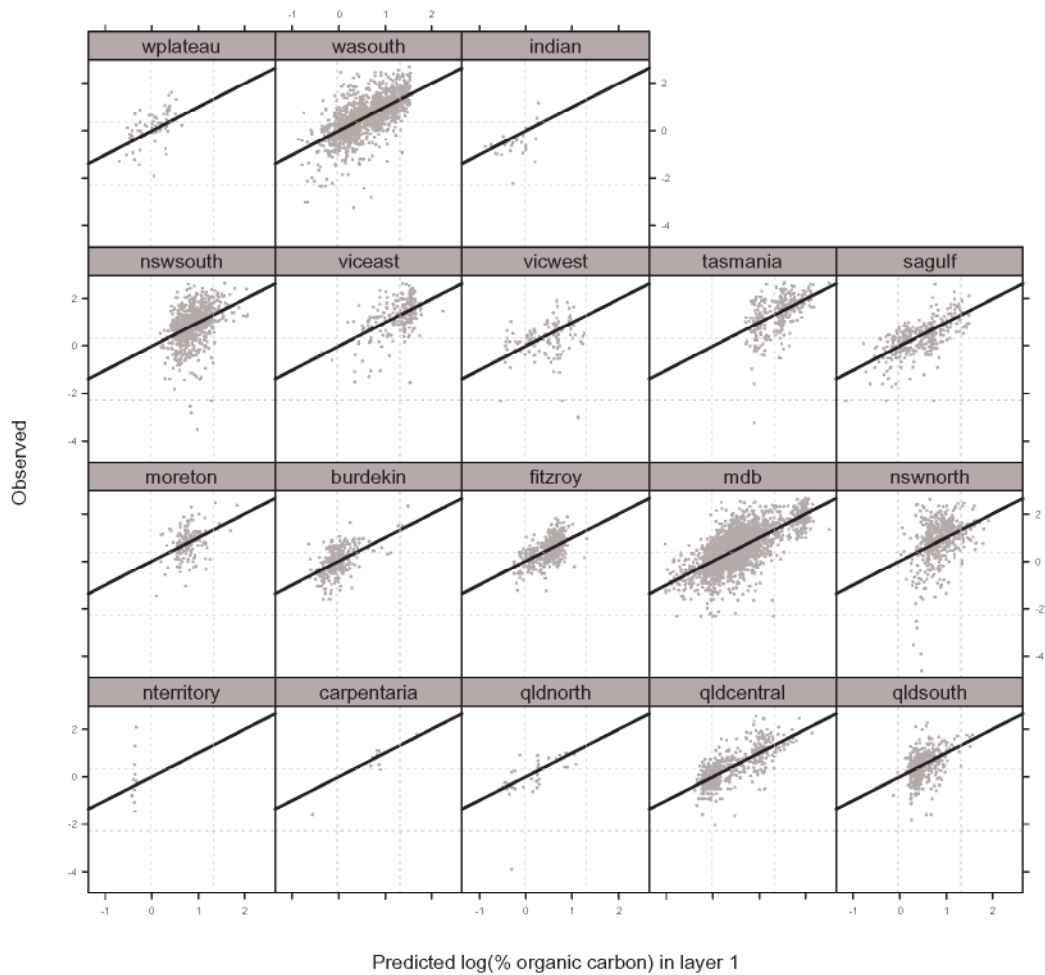


Figure 32: Observed versus predicted plots by region.

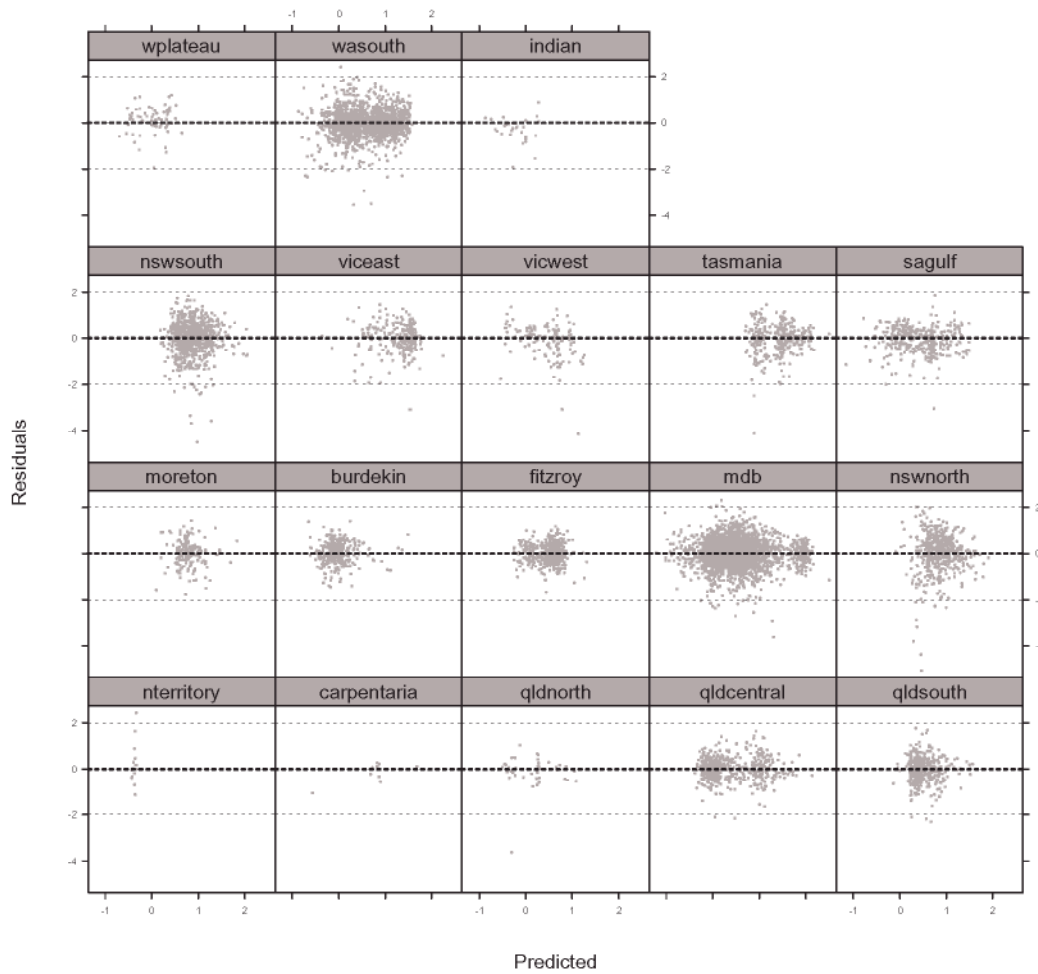


Figure 33: Residual plots by region.

The 29 rules from the final Cubist model were applied to ASRIS extent to generate a map of layer 1 organic carbon predictions. This map can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: elevation, relief, relative elevation, MSS band 2, annual mean moisture index, maximum temperature of the warmest period, annual precipitation and the annual mean radiation. This certainty surface can be seen at www.nlwra.gov.au/data.

5.2 Layer 2 organic carbon

The data as summarized by State/CSIRO and method are presented in Tables 25 and 26.

State	min	q10	q25	q50	q75	q90	max	N
NSW	0.02	0.19	0.30	0.50	0.89	1.58	13.10	1925
QLD	0.01	0.13	0.24	0.46	0.85	1.54	5.94	1068
SA	0.02	0.17	0.30	0.56	0.85	1.20	3.40	335
TAS	0.10	0.43	0.74	1.20	2.00	3.40	7.80	320
VIC	0.00	0.25	0.35	0.48	0.70	1.13	3.64	320
WA	0.00	0.11	0.18	0.32	0.59	1.04	5.59	794
CSIRO	0.06	0.20	0.30	0.50	1.15	1.96	7.26	666

Table 25: Distribution by State/CSIRO.

Method	min	q10	q25	q50	q75	q90	max	N
6A1	0.00	0.19	0.30	0.52	0.90	1.50	6.34	1078
6A1.UC	0.00	0.12	0.22	0.40	0.76	1.49	6.80	1095
6B1	0.02	0.18	0.29	0.49	0.92	1.60	13.10	1807
6B2	0.02	0.26	0.50	0.90	1.51	2.47	7.26	398
6B3	0.03	0.13	0.23	0.38	0.63	1.02	3.78	856
6.DC	0.10	0.39	0.60	1.10	2.00	3.40	7.80	194

Table 26: Distribution by organic carbon method.

Histograms of organic carbon in layer 2 are given in Figure 34. The range of the histograms is clipped to a maximum of 6 for illustration purposes. It is clear that the distributions are strongly right-skewed. Similarly to layer 1, a natural log transform was found to be beneficial to the modelling.

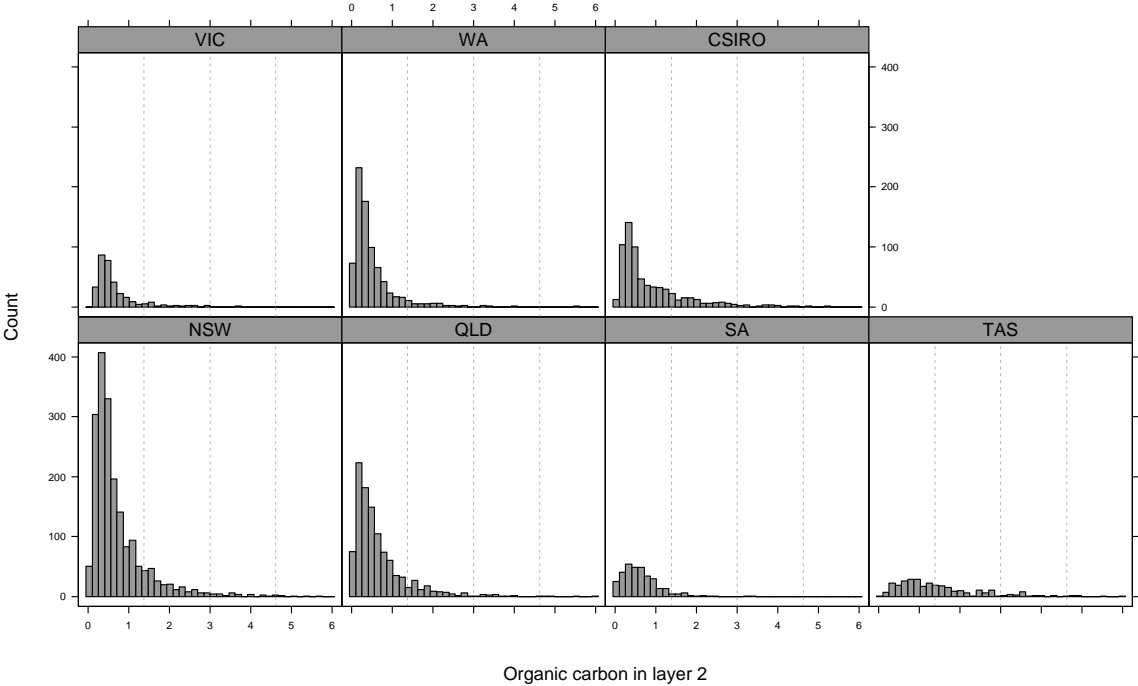


Figure 34: Histograms of layer 2 % organic carbon by State.

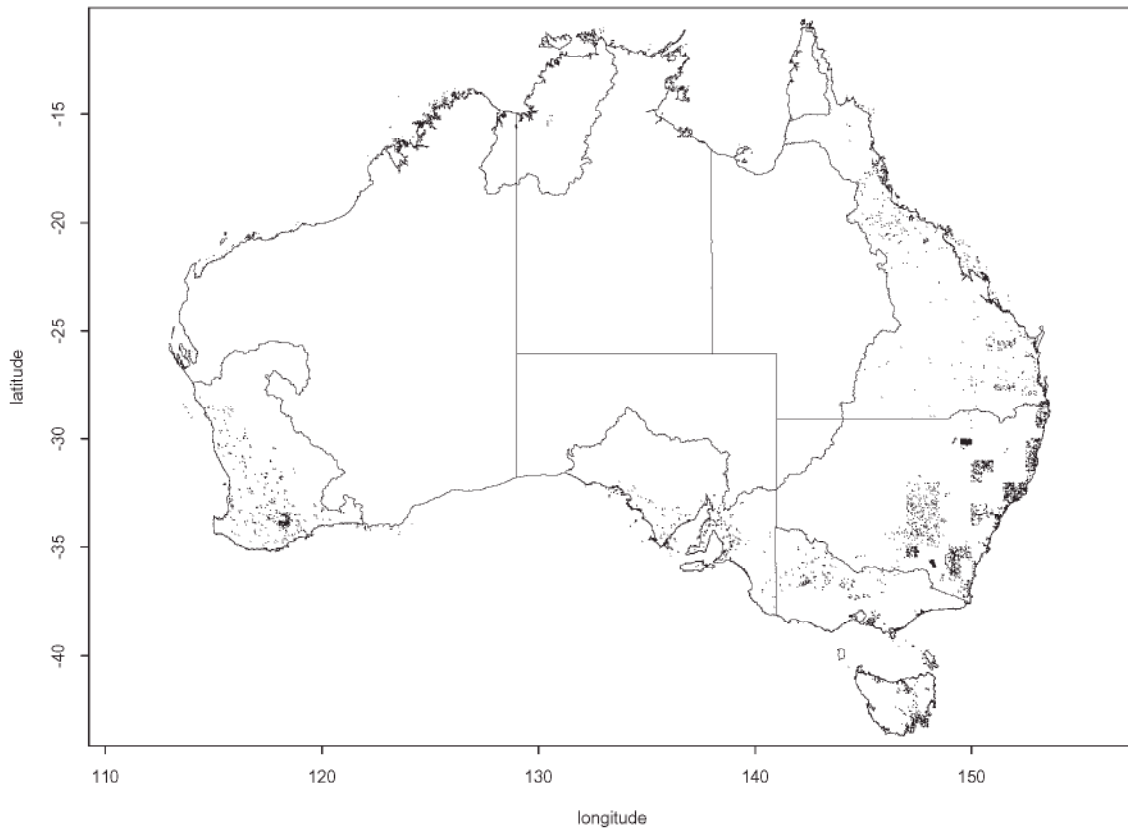


Figure 35: Locations of layer 2 % organic carbon observations.

The locations of the 5100 organic carbon observations used in the modelling are given in Figure 35.

A Cubist piecewise linear was fitted to the natural logarithm of these data. 31 variables were used: 11 climatic, 3 MSS, 13 terrain, lithology, landuse, ASC and the predicted layer 1 organic carbon surface. The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.24	0.77	0.59	0.84	0.50

Table 27: $\log(\text{organic carbon})$ in layer 2 model diagnostics on test data set.

The model is not as strong as layer 1 organic carbon. All quantities in Table 27 are poorer than their layer 1 equivalents. The overall performance on the test data is summarized graphically in Figure 36. There is a clear tendency to over-estimate the low $\log(\text{organic carbon})$ values. The residuals appear to be well-distributed when plotted against the fitted values. The quantile plot and histogram of these residuals are given in Figure 37.

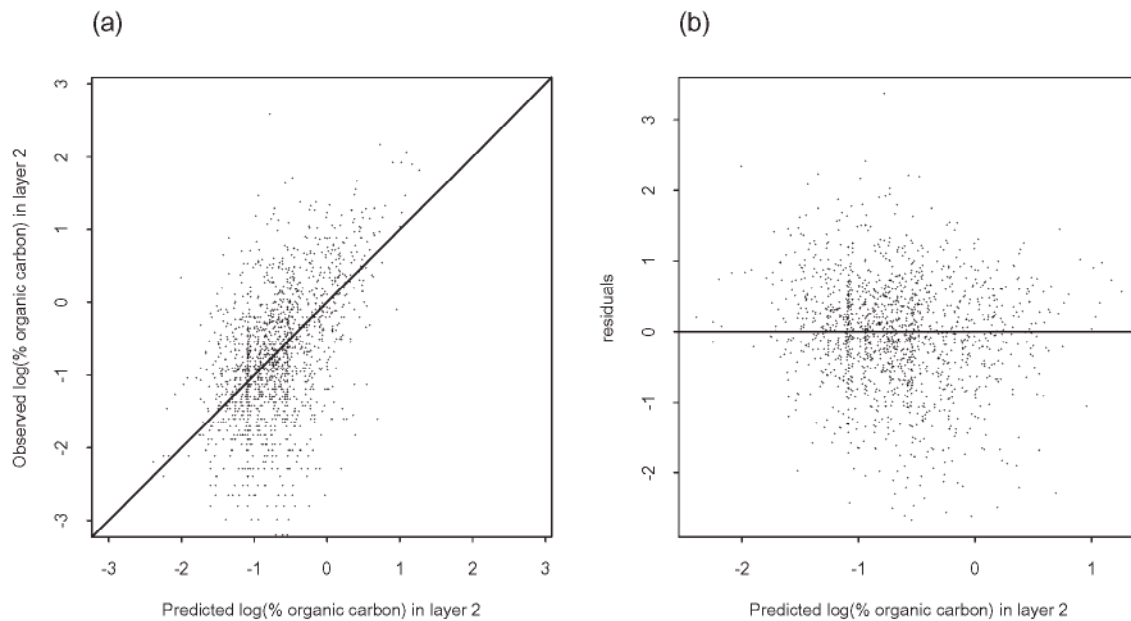


Figure 36: (a) Observed versus predicted and (b) residual plot for $\log(\% \text{ organic carbon})$ in layer 2 model (test data only).

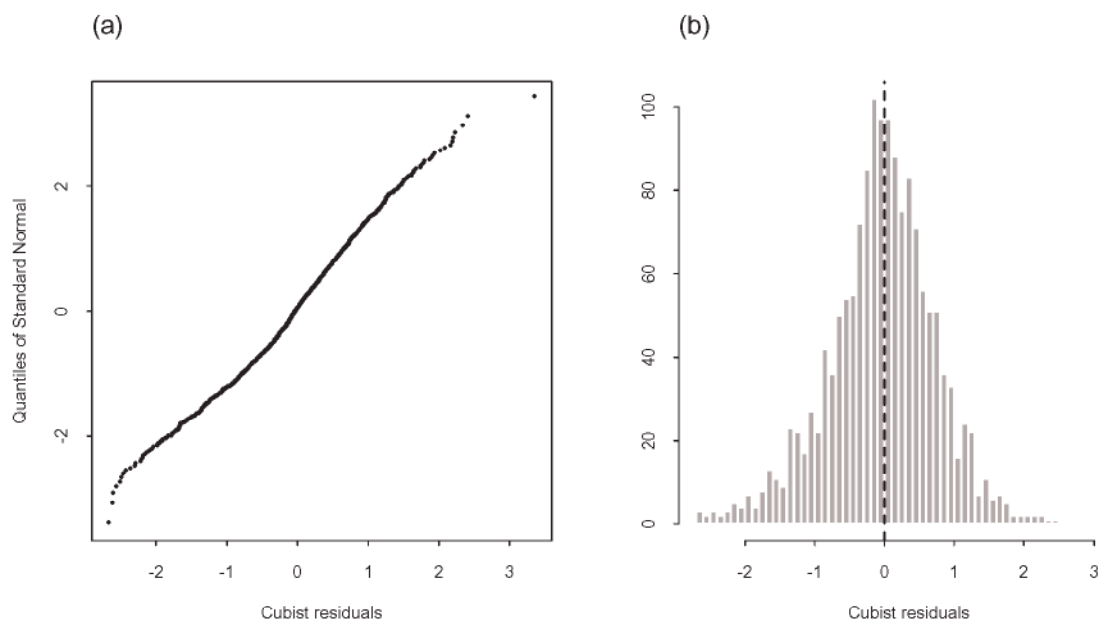


Figure 37: (a) Quantile plot and (b) histogram of Cubist residuals for $\log(\% \text{ organic carbon})$ in layer 2 model (test data).

The model was then refitted using the same model options and variables on all 5100 observations. 19 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 0.40; relative error 0.74; correlation 0.54).

The state and region-wise performance are summarized in Tables 28 and 29.

Figure 36 can be decomposed into the 18 regions making up the ASRIS extent. This leads directly to Figure 38 and Figure 39 and enables some spatial assessment of performance.

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
NSW	1847	0.45	0.86	0.58	0.76
QLD	970	0.65	0.74	0.55	0.72
SA	328	0.38	0.94	0.59	0.75
TAS	253	0.61	0.78	0.47	0.58
VIC	303	0.35	0.88	0.41	0.53
WA	754	0.47	0.87	0.60	0.76
CSIRO	645	0.70	0.63	0.47	0.62

Table 28: Performance of organic carbon in layer 2 model by state/CSIRO.

Region	N	rank correlation	relative error	average error	approximate RMSE
nterritory	18	0.02	1.15	0.54	0.70
carpentaria	7	0.70	1.06	0.36	0.50
qldnorth	10	0.83	0.85	0.50	0.58
qldcentral	245	0.60	0.76	0.55	0.74
qldsouth	113	0.71	0.65	0.52	0.70
moreton	52	0.50	0.85	0.75	0.98
burdekin	136	0.50	0.89	0.47	0.58
fitzroy	8	0.36	1.42	0.41	0.51
mdb	2021	0.59	0.75	0.50	0.65
nswnorth	429	0.37	0.91	0.64	0.83
nswsouth	604	0.45	0.86	0.59	0.78
viceast	173	0.40	0.89	0.64	0.78
vicwest	31	0.04	1.02	0.76	0.90
tasmania	253	0.61	0.78	0.47	0.58
sagulf	226	0.40	0.92	0.53	0.68
wplateau	30	0.07	1.12	0.64	0.82
wasouth	711	0.46	0.87	0.60	0.76
indian	33	0.44	0.94	0.70	0.87

Table 29: Performance of organic carbon in layer 2 model by region.

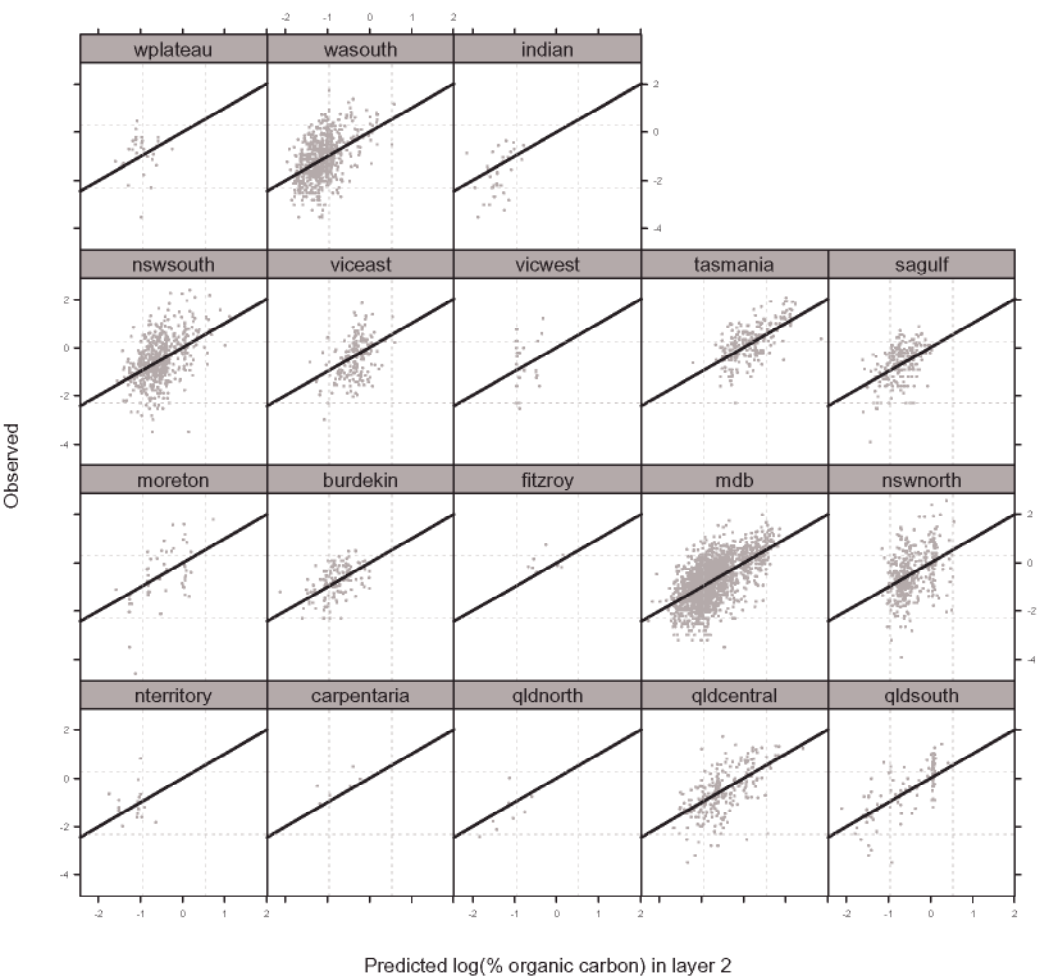


Figure 38: Observed versus predicted plots by region.

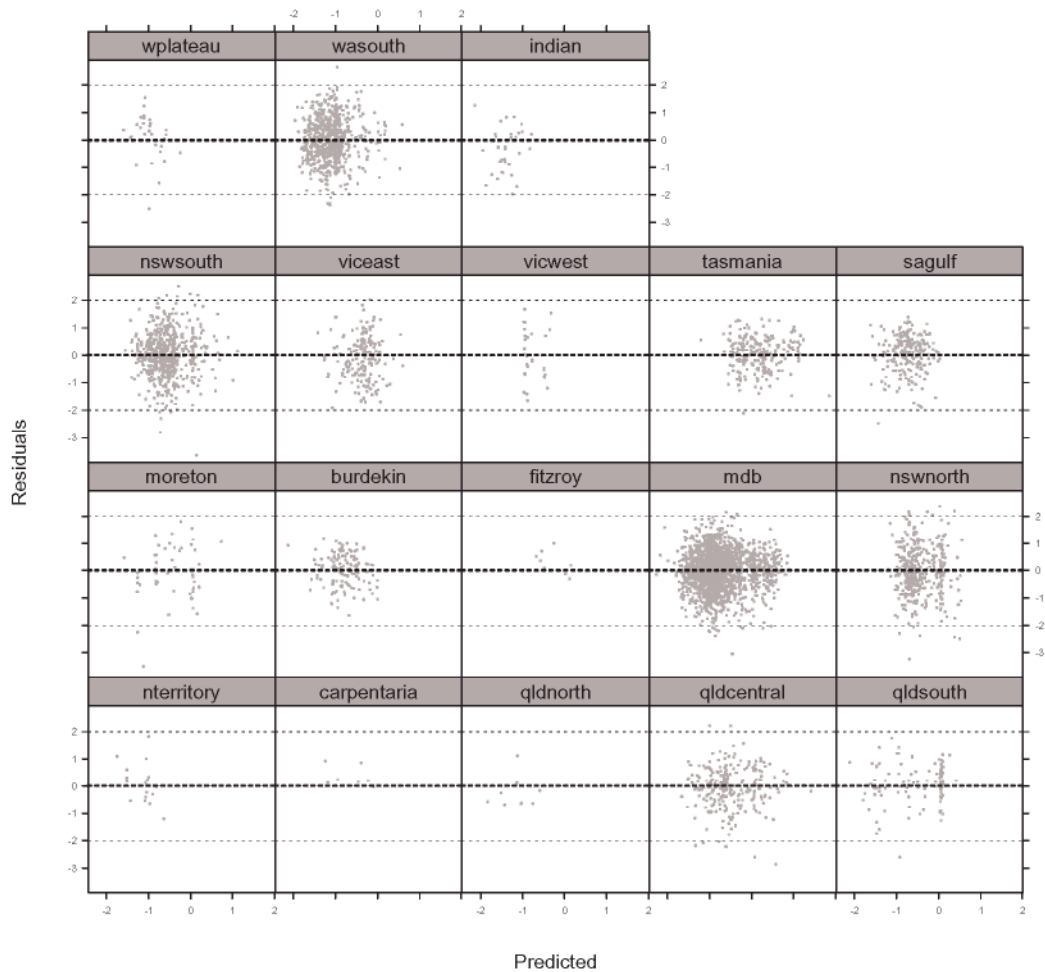


Figure 39: Residual plots by region.

The model performance is stronger in southern/central Queensland, Tasmania and the Murray-Darling basin. Western Australia and South Australia are fair. The predictions in coastal Victoria and New South Wales are weakly supported.

The 19 rules from the final Cubist model were applied to ASRIS extent to generate a map of layer 2 organic carbon predictions. These predictions can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: elevation, relief, relative elevation, MSS band 2, annual mean moisture index, maximum temperature of the warmest period, annual precipitation and the annual mean radiation. This certainty surface is available at www.nlwra.gov.au/data.

6 NITROGEN

6.1 *Layer 1 total nitrogen*

Layer 1 nitrogen was predicted indirectly by appealing to the relationship between nitrogen and organic carbon. While a continental model for total nitrogen could be derived analogously to other soil properties, the paucity of total nitrogen observations in New South Wales and South Australia affected its reliability. An indirect approach was favoured because of the strength of the observed relationship between nitrogen and organic carbon, and the fact that the larger presence of organic carbon in the ASRIS database would make for a stronger organic carbon point model, and thus a potentially more stable continental nitrogen model.

There were a number of assessment methods used for both nitrogen and organic carbon. For total nitrogen, methods 7A1, 7A2 and 7A5 were all considered. Descriptions of these methods can be found in Rayment & Higginson (1992). The organic carbon methods used, namely 6A1, 6A1.UC, 6B1, 6B2, 6B3 and 6.DC, were described in Section 5

There were 4746 observations with both layer 1 total nitrogen and organic carbon assessments. The locations of these observations are given in Figure 40. The sparsity of observations in New South Wales and South Australia arises because there were very few total nitrogen measurements made in either State.

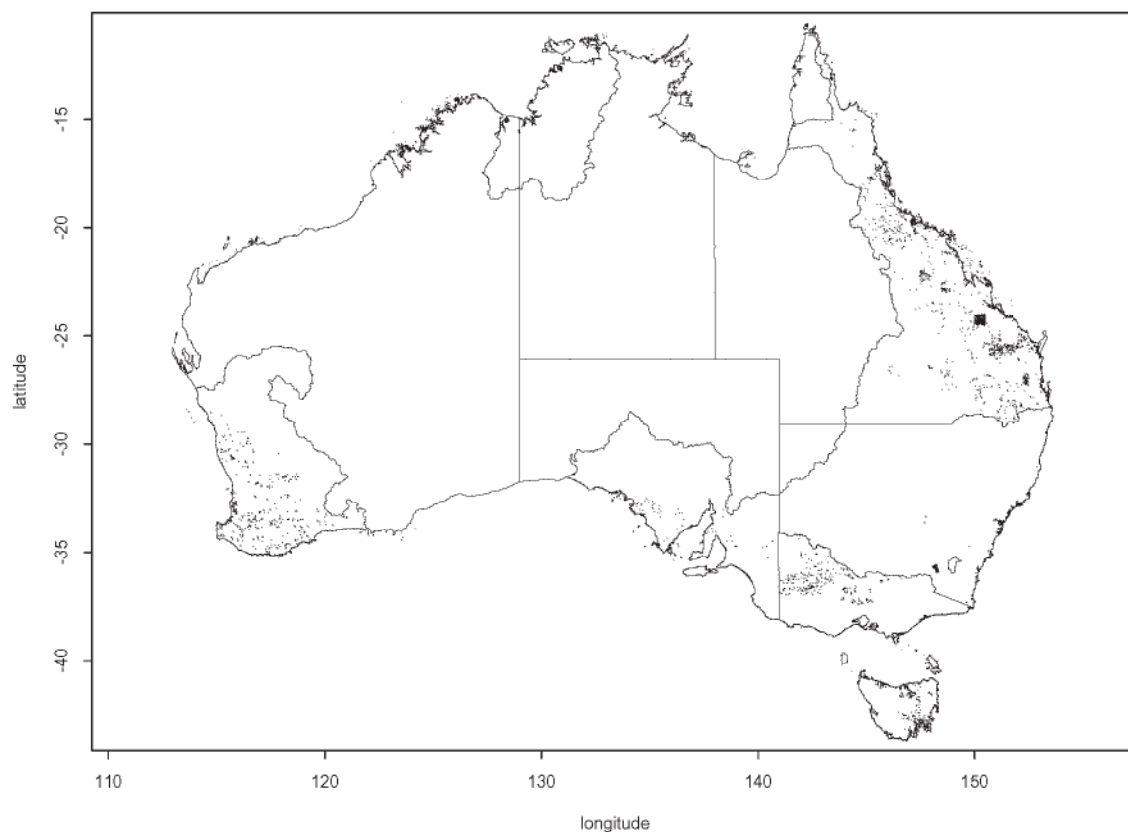


Figure 40: Locations with both layer 1 total nitrogen and organic carbon.

The respective nitrogen and organic carbon values are displayed in Figure 41 along with the regression line from the linear regression on the log-scale, namely

$$\log_e(\text{nitrogen}) = -2.6589 + 0.8761 \log_e(\text{organic carbon}).$$

Figure 41 clearly indicates a strong relationship between nitrogen and carbon, though there is some suggestion of over-predicting nitrogen at the low end (RMSE = 0.42 on the log scale, $R^2 = 0.75$). Note, that the data here are displayed on a \log_{10} scale here purely to take advantage of the log-scale plotting.

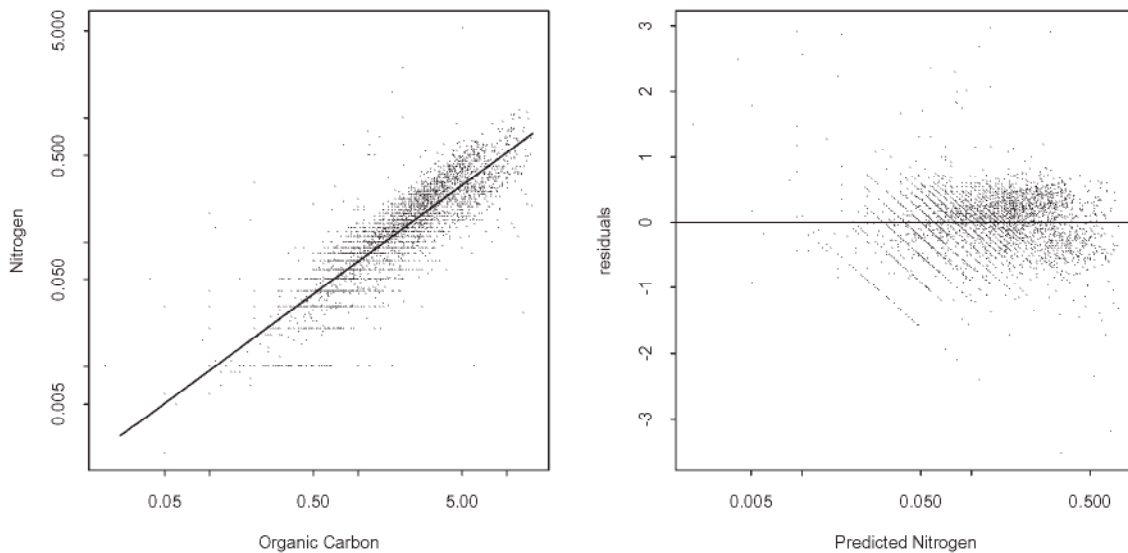


Figure 41: Nitrogen versus organic carbon (log-scale).

A predicted $\log(\text{nitrogen})$ surface was derived by applying the regression equation to the modelled surface for $\log(\text{organic carbon})$ in layer 1. The structure inherent in the $\log(\text{nitrogen})$ surface was thus necessarily identical to that of $\log(\text{organic carbon})$. This prediction surface can be viewed at www.nlwra.gov.au/data.

The certainty in this modelled surface was heavily dependent on both the strength of the layer 1 organic model and the strength of the relationship between nitrogen and organic carbon. The approach taken here was to use the certainty surface for organic carbon as that captured the first component and then downweight it to reflect the fact that nitrogen was not estimated directly, but rather through the derived relationship with organic carbon.

7 PHOSPHORUS

The total phosphorus measurements required a considerable amount of cleaning. In particular, there were unit inconsistencies with some measurements recorded in parts per million (ppm) and others in terms of %. For the most part the unit was consistent within the agencies which made it easier to report all measurements as % (after the necessary conversion for agencies with ppm). Where both ppm and %'s were used within an agency, individual decisions were made on the measurements. Some observations were deleted as their units could not be ascertained.

7.1 Layer 1 total phosphorus

Table 30 provides an indication of the distributions of total phosphorus across the States/CSIRO. The counts are clearly dominated by Queensland and Western Australia.

State	min	q10	q25	q50	q75	q90	max	N
NSW	0.010	0.010	0.020	0.020	0.050	0.082	0.150	45
QLD	0.001	0.009	0.015	0.030	0.050	0.100	0.880	2488
SA	0.001	0.004	0.009	0.010	0.020	0.040	0.280	453
TAS	0.001	0.004	0.009	0.020	0.052	0.109	0.368	511
VIC	0.001	0.002	0.004	0.006	0.008	0.010	0.010	22
WA	0.000	0.001	0.002	0.003	0.004	0.010	0.920	4563
CSIRO	0.002	0.004	0.010	0.030	0.061	0.151	0.474	321

Table 30: Distribution by State/CSIRO.

There are a large number of methods used to assess total phosphorus. Those considered here are given in Table 31. Method 9A.NR represents an assessment of total phosphorus by an unknown method.

Method	min	q10	q25	q50	q75	q90	max	N
9A1	0.001	0.008	0.016	0.030	0.054	0.100	0.880	2001
9A3	0.001	0.003	0.007	0.018	0.038	0.090	0.474	678
9A.HCL	0.000	0.002	0.006	0.010	0.030	0.080	0.920	877
9A.HCLP2O5	0.007	0.009	0.010	0.010	0.020	0.040	0.040	11
9A.HF+	0.006	0.008	0.018	0.022	0.030	0.046	0.205	38
9A.NR	0.000	0.001	0.002	0.003	0.006	0.020	0.780	4798

Table 31: Distribution by total phosphorus method.

Histograms of layer 1 phosphorus by State are given in Figure 42. These highlight the strong presence of data from Queensland and Western Australia. There is some indication of digit preference in Queensland as evidenced by the clustering about certain values.

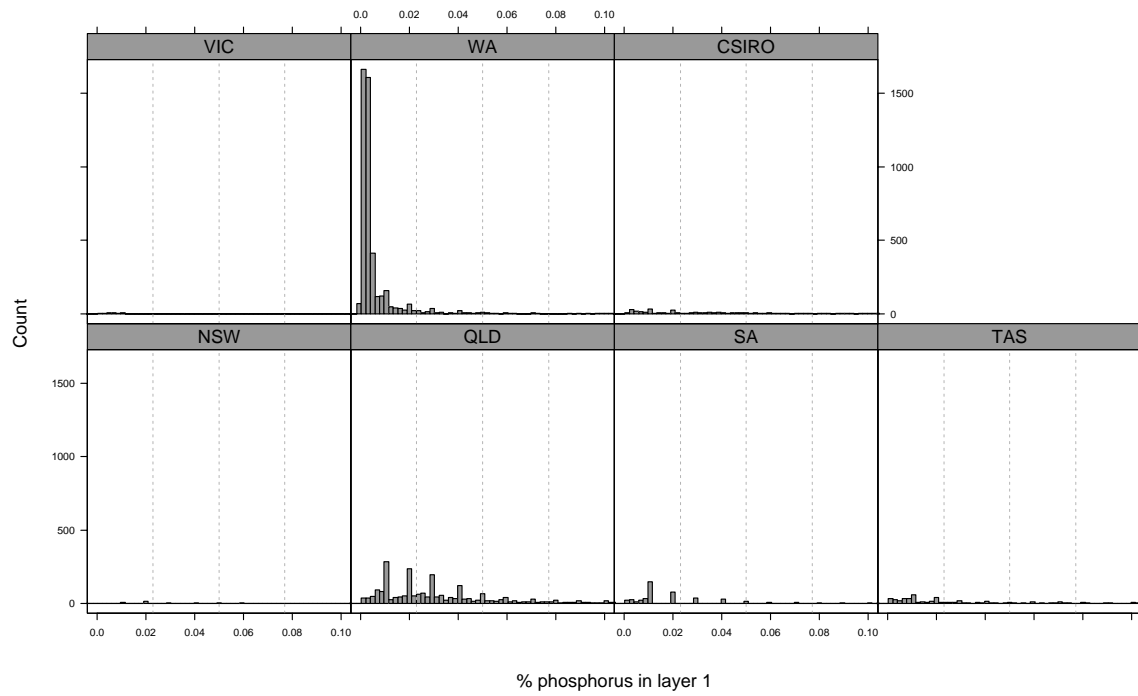


Figure 42: Histograms of layer 1 total phosphorus by State.

The locations of the 7377 observations used in the model that fall inside the ASRIS extent are given in Figure 43. This again highlights the paucity of total phosphorus data in New South Wales and Victoria.

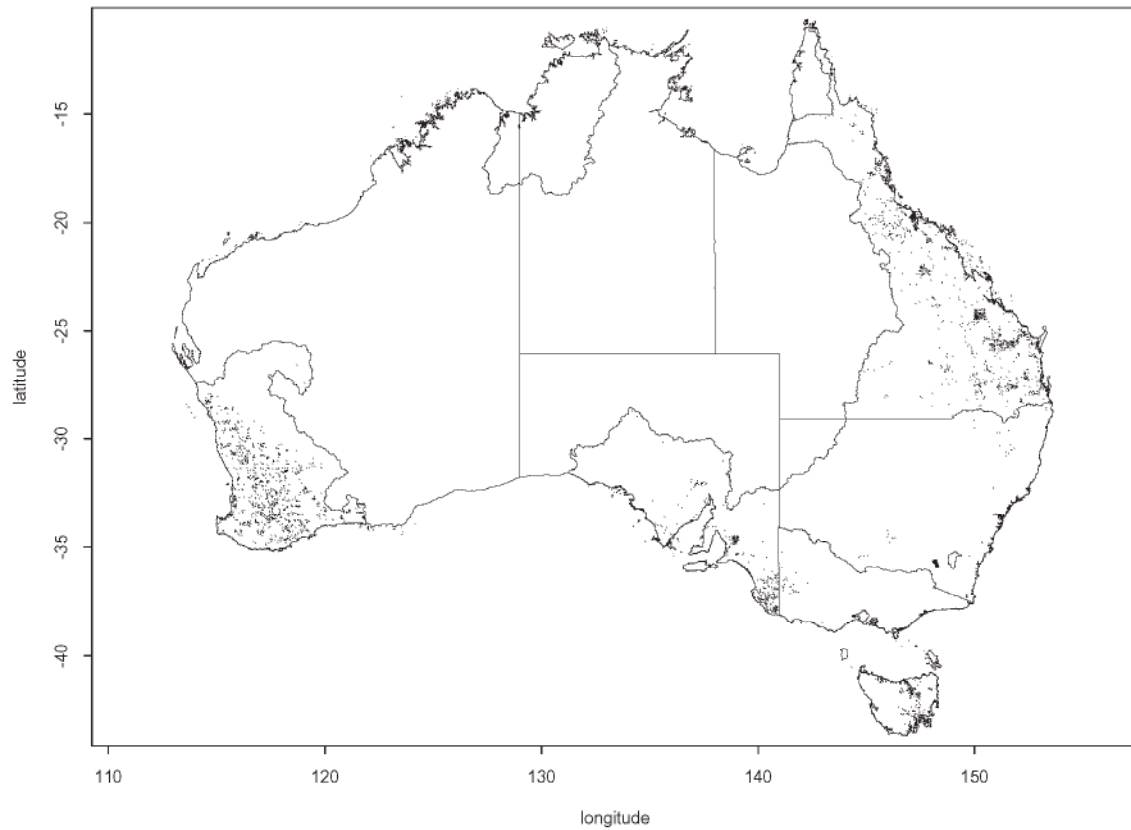


Figure 43: Locations of layer 1 total phosphorus observations.

A *Cubist* piecewise linear model was fitted to these data. 30 variables were used: 10 climatic, 3 MSS, 14 terrain, lithology, ASC and predicted phosphorus from a polygon model (via a PPF look-up table). The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.62	0.92	0.68	0.54	0.79

Table 32: log(phosphorus) in layer 1 model diagnostics on test data set.

The overall performance on the test data is summarized graphically in Figure 44. There is some suggestion of clustering in the residuals. Overall the residuals appear fairly well distributed on the log-scale. The quantile plot and histogram of residuals are given in Figure 45.

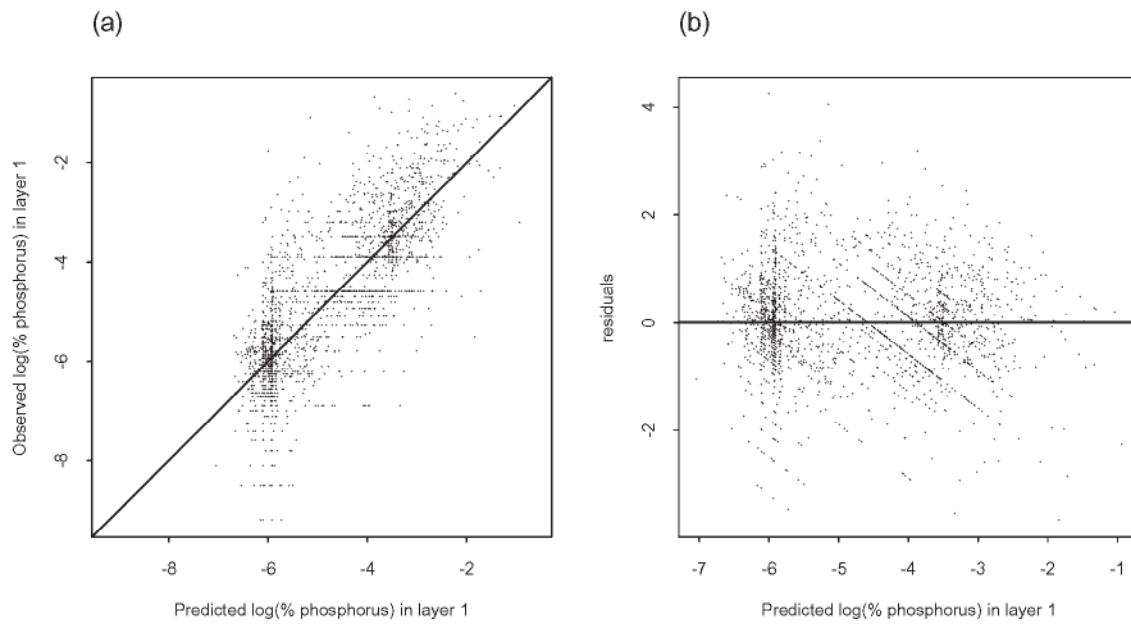


Figure 44: (a) Observed versus predicted and (b) residual plot for log(% phosphorus) in layer 1 model (test data only).

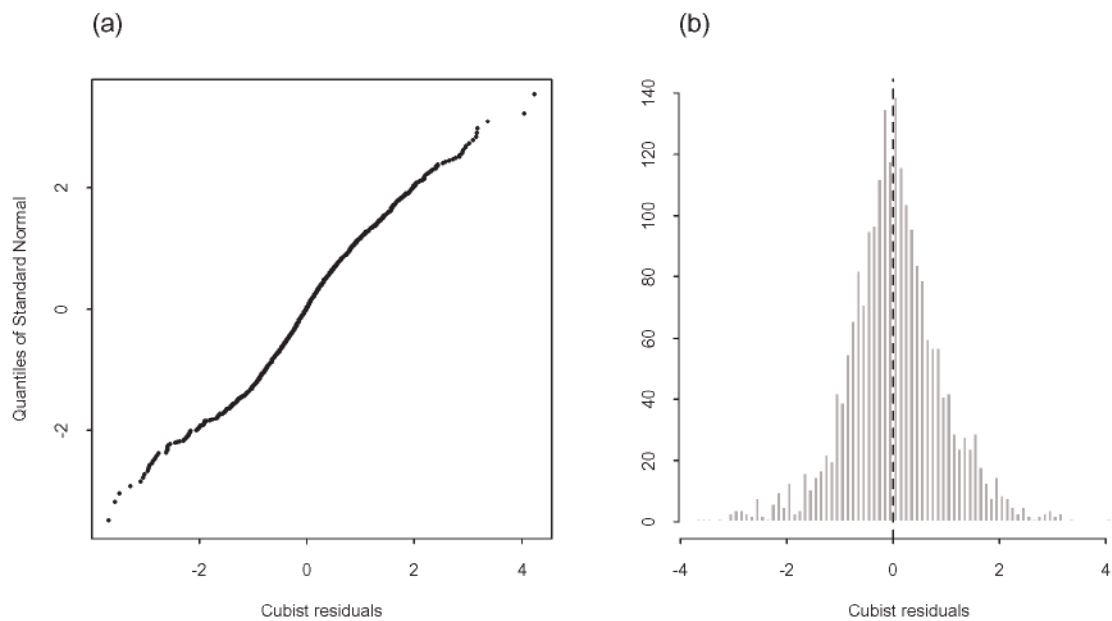


Figure 45: (a) Quantile plot and (b) histogram of Cubist residuals for log(% phosphorus) in layer 1 model (test data).

The model was then refitted using the same model options and variables on all 7377 observations. 18 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 0.63; relative error 0.50; correlation 0.82).

The state and region-wise performance are summarized in Tables 33 and 34.

Figure 44 can be decomposed into the 18 regions making up the ASRIS extent. This leads directly to Figure 46 and Figure 47 and enables some spatial assessment of performance.

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
NSW	42	0.39	1.04	0.64	0.88
QLD	2231	0.60	0.76	0.58	0.76
SA	429	0.37	0.91	0.68	0.90
TAS	423	0.59	0.70	0.66	0.87
VIC	21	0.14	1.27	0.58	0.78
WA	3931	0.42	0.86	0.63	0.87
CSIRO	250	0.81	0.50	0.56	0.69

Table 33: Performance of phosphorus in layer 1 model by state/CSIRO.

Region	N	rank correlation	relative error	average error	approximate RMSE
nterritory	10	-0.66	1.33	0.25	0.31
carpentaria	7	0.89	0.74	0.99	1.06
qldnorth	61	0.35	0.86	0.82	1.11
qldcentral	577	0.57	0.74	0.61	0.79
qldsouth	493	0.54	0.78	0.58	0.79
moreton	117	0.58	0.77	0.73	0.94
burdekin	436	0.57	0.79	0.58	0.74
fitzroy	281	0.50	0.88	0.46	0.56
mdb	515	0.67	0.76	0.54	0.69
nswnorth	8	0.50	1.42	1.20	1.46
nswsouth	6	0.18	2.58	0.87	1.05
viceast	58	0.27	1.15	0.44	0.53
vicwest	224	0.48	0.83	0.74	0.95
tasmania	423	0.59	0.70	0.66	0.87
sagulf	158	0.23	1.06	0.59	0.82
wplateau	32	0.25	1.13	0.56	0.80
wasouth	3576	0.42	0.86	0.64	0.89
indian	345	0.31	0.93	0.51	0.69

Table 34: Performance of phosphorus in layer 1 model by region.

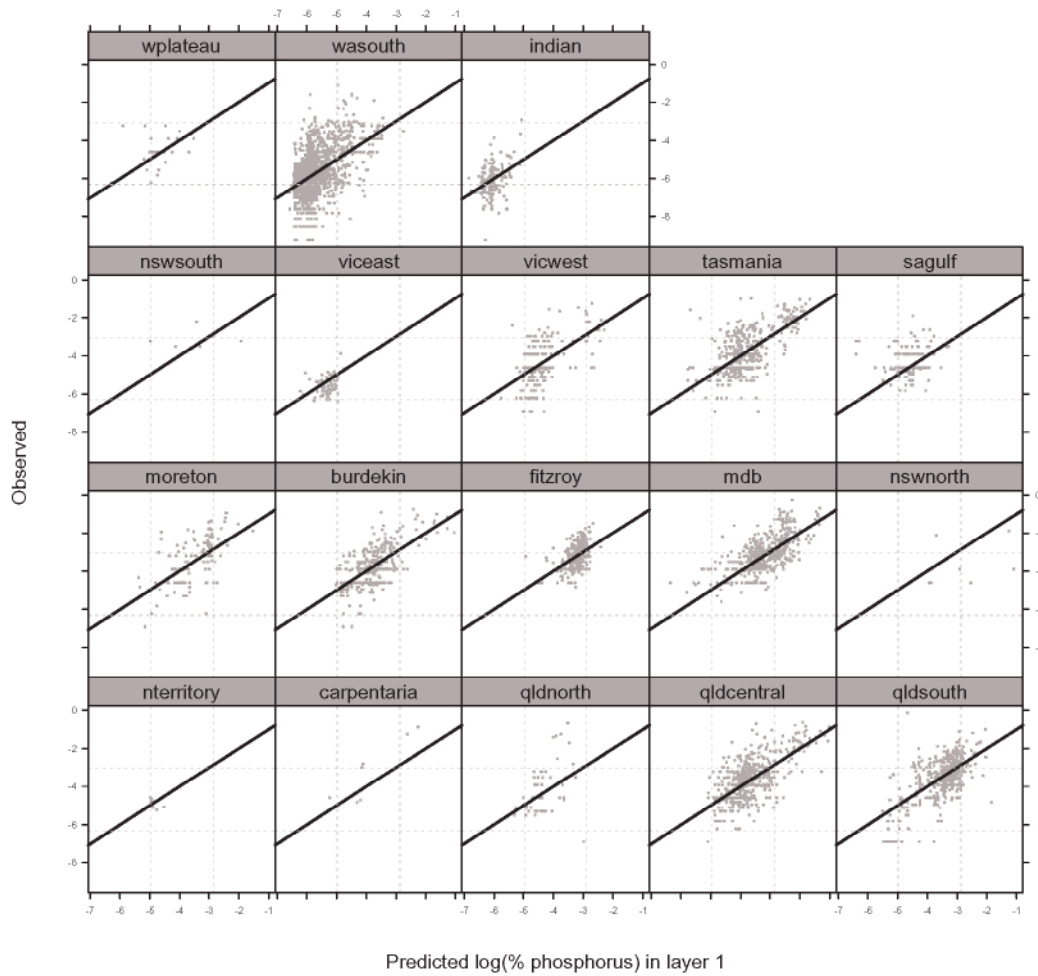


Figure 46: Observed versus predicted plots by region.

The performance in Queensland, the Murray-Darling basin and Tasmania are encouraging. In Western Australia, from where a large proportion of the data derive, there is a concentration of low phosphorus values and the model does not appear to do well at predicting anything else. This explains some of the clustering in the Figure 44. The predictions in South Australia are not well supported. There are too few observations in New South Wales to draw any conclusions.

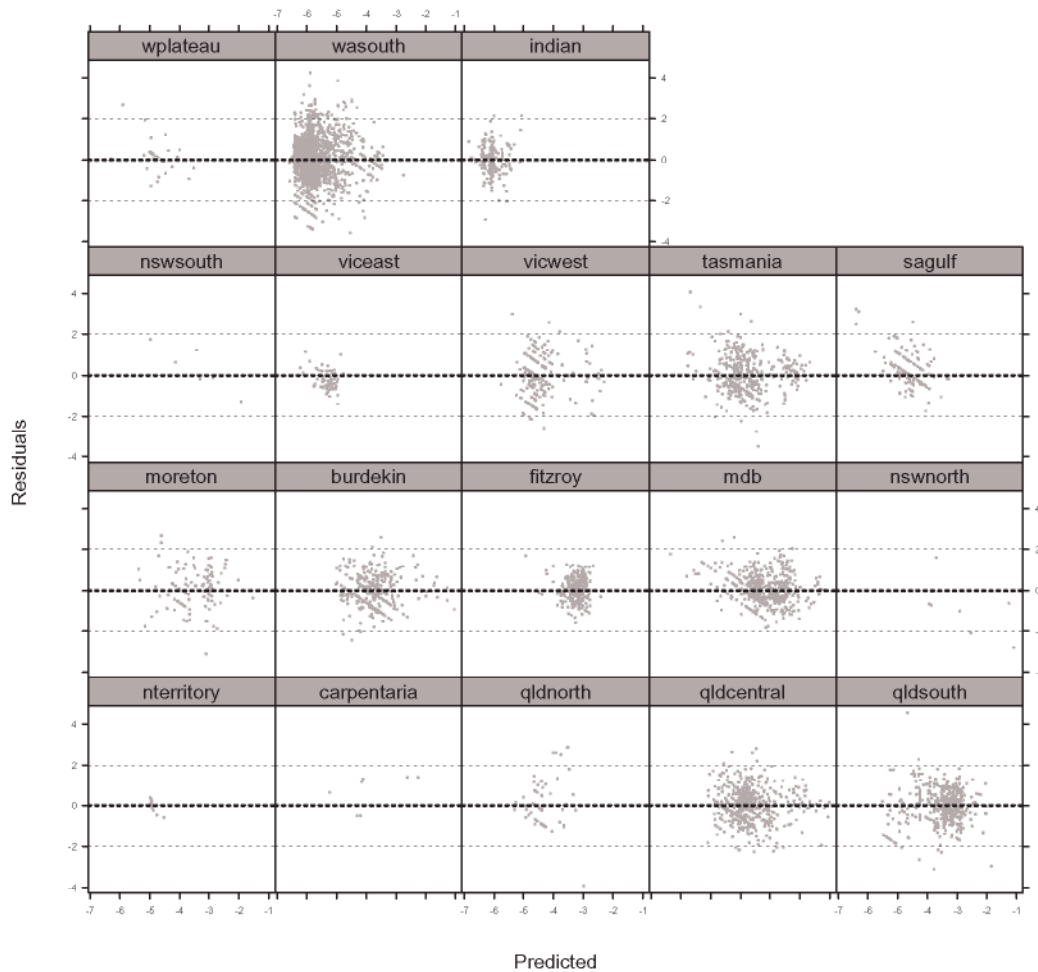


Figure 47: Residual plots by region.

The 18 rules from the final Cubist model were applied to ASRIS extent to generate a map of layer 1 total phosphorus predictions. These predictions can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: elevation, relief, relative elevation, MSS band 3, moisture index seasonality, lowest period radiation, precipitation of the wettest period and the minimum temperature of the coldest period. The environmental representativeness component is of particular interest here. It reflects the fact that while we do not have many observations in Victoria or New South Wales, it is still reasonable to make predictions into those States because a large part of the environmental space is similar to areas in which we do have points. This certainty surface can be seen at www.nlwra.gov.au/data.

7.2 Layer 1 extractable phosphorus

Extractable phosphorus in parts per million (ppm) was considered for New South Wales and Victoria, largely because the total phosphorus was very poorly represented in the database for both States. These data are summarised in Table 35.

State	min	q10	q25	q50	q75	q90	max	N
NSW	1.0	2.0	3.0	5.00	11.00	25.00	150.0	2108
VIC	1.2	2.2	3.1	6.55	12.55	24.35	73.8	122

Table 35: Distribution by State/CSIRO (ppm).

3 large outliers were truncated back to 150 ppm. Some small readings were deleted as it could not be ascertained whether the units were particularly small ppm readings or percentages.

Three methods were considered for extractable P, namely 9B.9C, 9C2 and 9E1. Details of the methods can be found in Rayment & Higginson (1992). The 2124 extractable phosphorus observations that were available, fell within the ASRIS extent and had all environmental predictors available are given in Figure 48.

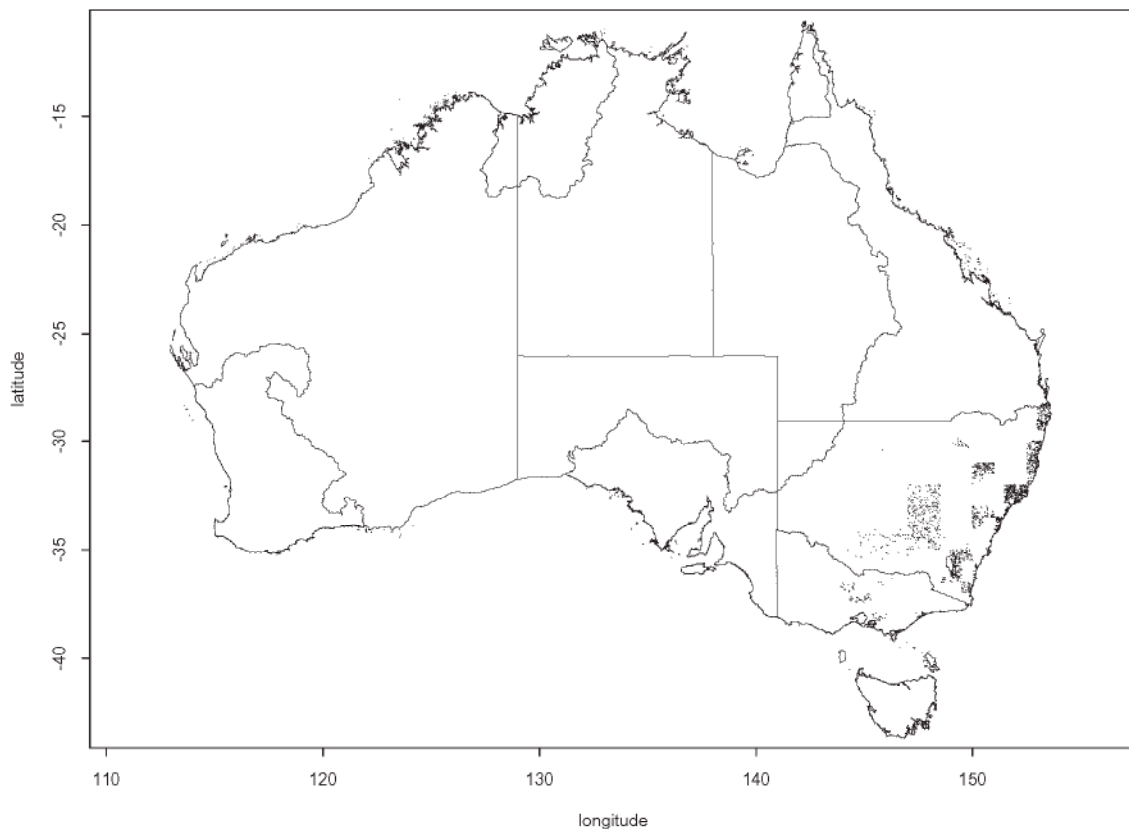


Figure 48: Locations of layer 1 extractable phosphorus observations used in the modelling.

A *Cubist* piecewise linear model was fitted to these data. 30 variables were used: 10 climatic, 3 MSS, 14 terrain, lithology, ASC and predicted phosphorus from a polygon model (via a PPF look-up table). The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.35	0.85	0.67	0.78	0.59

Table 36: log(phosphorus) in layer 1 model diagnostics on test data set

The overall performance on the test data is summarized graphically in Figure 49. The predictive power is fairly weak. There is a clear tendency to over-predict low extractable phosphorus values and under-predict higher extractable phosphorus.

The quantile plot and histogram of residuals are given in Figure 50.

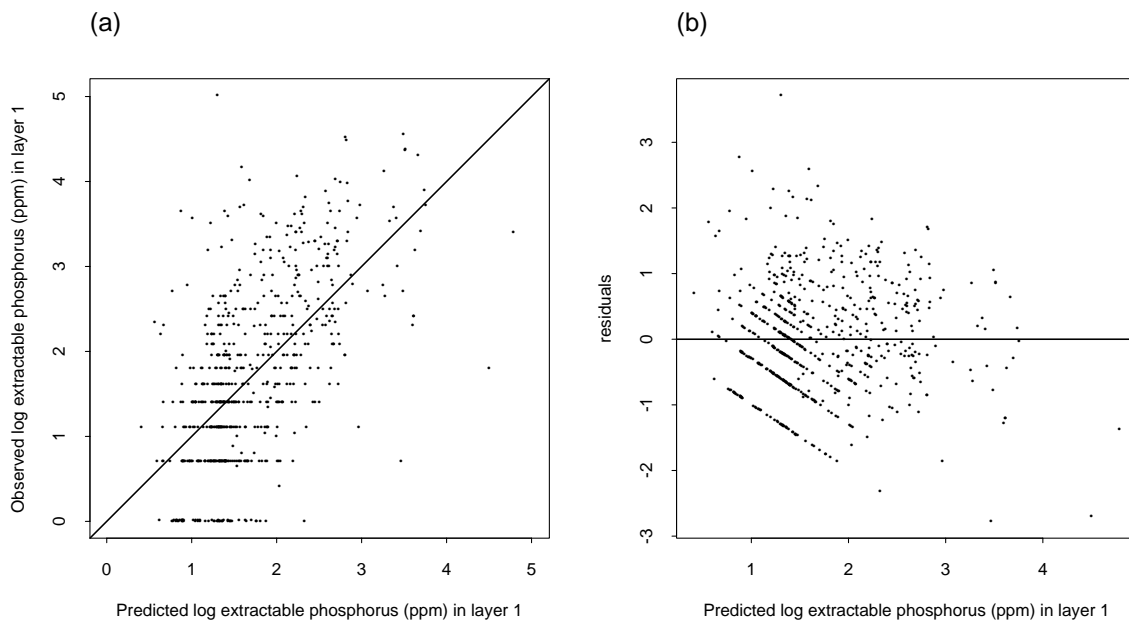


Figure 49: (a) Observed versus predicted and (b) residual plot for log extractable phosphorus (ppm) in layer 1 model (test data only).

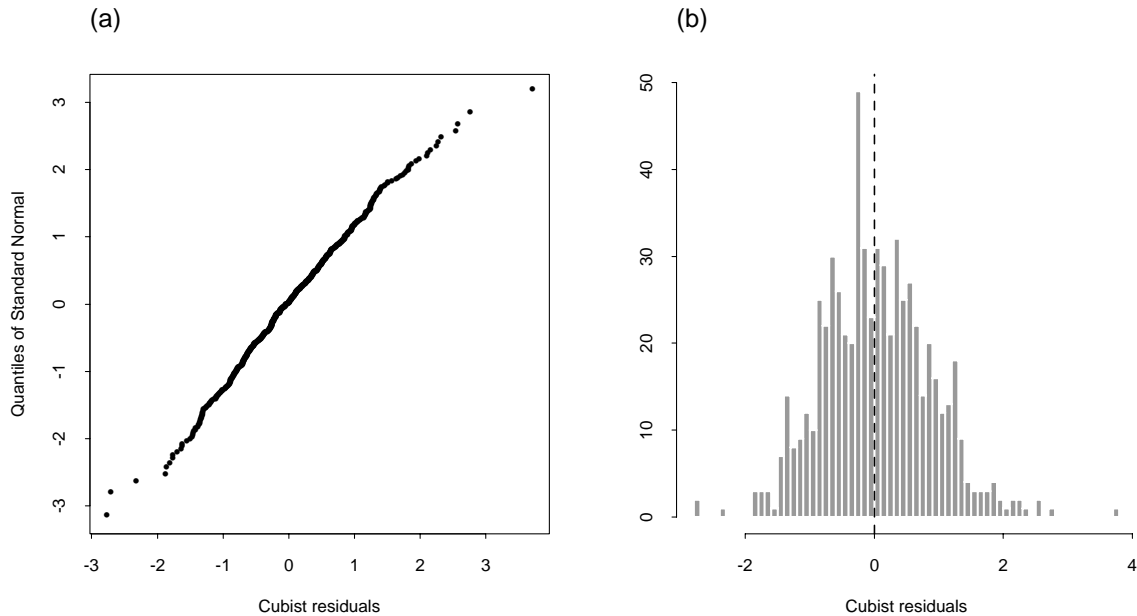


Figure 50: (a) Quantile plot and (b) histogram of Cubist residuals for log extractable phosphorus in layer 1 model (test data).

The model was then refitted using the same model options and variables on all 2124 observations. 18 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 0.63; relative error 0.75; correlation 0.62).

Table 37 breaks down the overall performance from the model fitted to all the data into that of the four ASRIS regions that comprise the New South Wales/Victoria observation set. These diagnostics suggest that performance is marginally stronger in the *mbd* and *viceast*.

Region	N	rank correlation	relative error	average error	approximate RMSE
mdb	929	0.59	0.77	0.65	0.83
nswnorth	445	0.44	0.88	0.59	0.77
nswsouth	615	0.38	0.89	0.58	0.75
viceast	135	0.64	0.72	0.64	0.80

Table 37: Performance of extractable phosphorus in layer 1 model by region.

The 18 rules from the final Cubist model were applied to those error regions in Figure 3 falling in New South Wales and Victoria to generate a map of layer 1 extractable phosphorus predictions. These predictions can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: elevation, relief, relative elevation, MSS band 3, moisture index seasonal-

ity, lowest period radiation, precipitation of the wettest period and the minimum temperature of the coldest period. This certainty surface is available at www.nlwra.gov.au/data.

8 CLAY

8.1 Layer 1 clay

The distribution of layer 1 % clay across the States/CSIRO is given in Table 38 and Figure 51. Inspection suggests that there are more sites with a higher clay content in Queensland and Victoria.

State	min	q10	q25	q50	q75	q90	max	N
NSW	0	4	8	13	23	39	84	2571
QLD	0	6	12	26	45	58	94	3098
SA	0	4	6	10	21	33	73	1288
TAS	0	3	7	16	30	51	78	584
VIC	0	8	13	21	36	50	81	514
WA	0	2	4	7	13	21	82	1654
CSIRO	1	8	13	20	36	54	75	813

Table 38: Distribution by State/CSIRO.

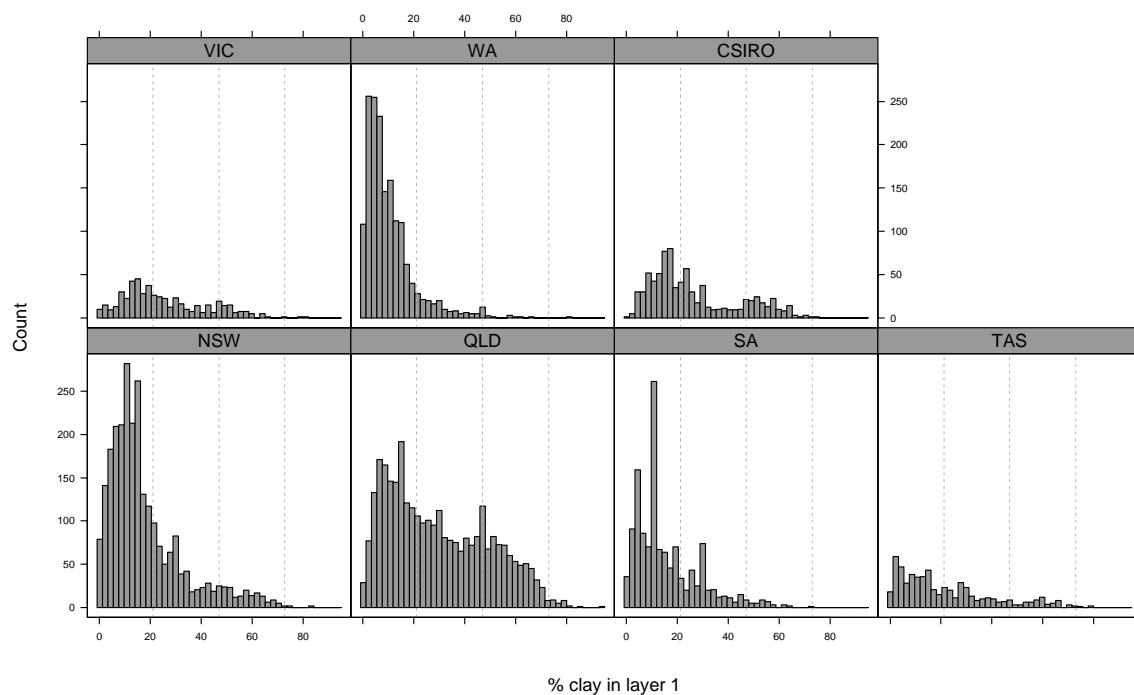


Figure 51: Histograms of layer 1 % clay by State.

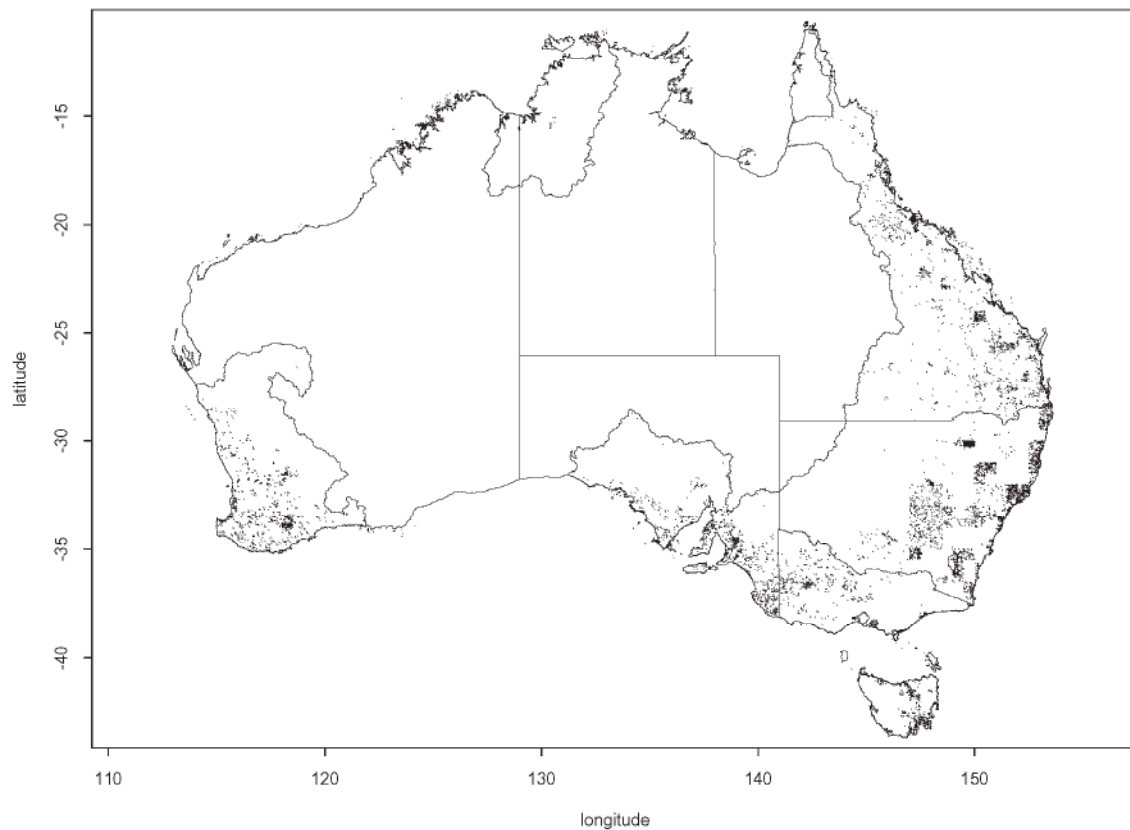


Figure 52: Locations of layer 1 % clay observations.

The locations of the 9750 layer 1 % clay observations used in the modelling are given in Figure 52.

A Cubist piecewise linear model was fitted to the square root of these data as this transformation was found to reduce the skewness and stabilize the variance. 33 variables were used: 11 climatic, 3 MSS, 15 terrain, lithology, landuse, ASC and predicted % clay from a polygon model (via PPF derived look-up table). The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.44	1.36	1.05	0.70	0.67

Table 39: sqrt(% clay) in layer 1 model diagnostics on test data set.

The overall performance on the test data is summarized in Figure 53. Layer 1 (square root) clay clearly exhibits predictive ability. There is however a large degree of scatter in these predictions. Moreover, there is some tendency to over-predict soils with low clay content and under-predict the higher clay content. While the residuals may not appear particularly well distributed about 0 upon first inspection, this impression might belie the relative point density. The quantile plot and histogram of residuals are given in Figure 54.

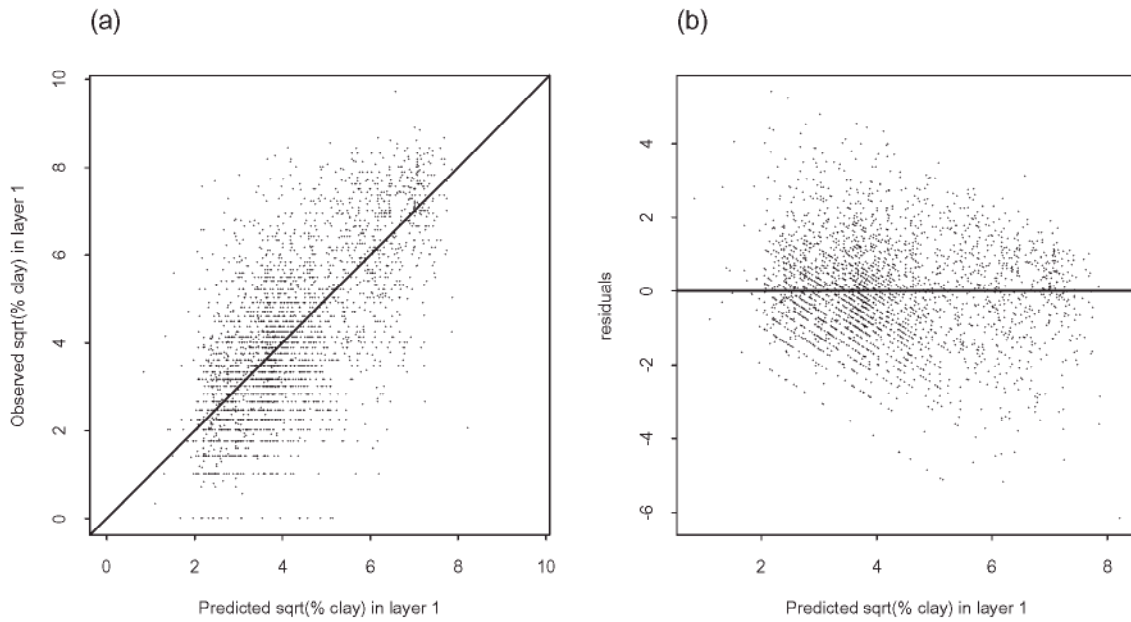


Figure 53: (a) Observed versus predicted and (b) residual plot for sqrt(% clay) in layer 1 model (test data only).

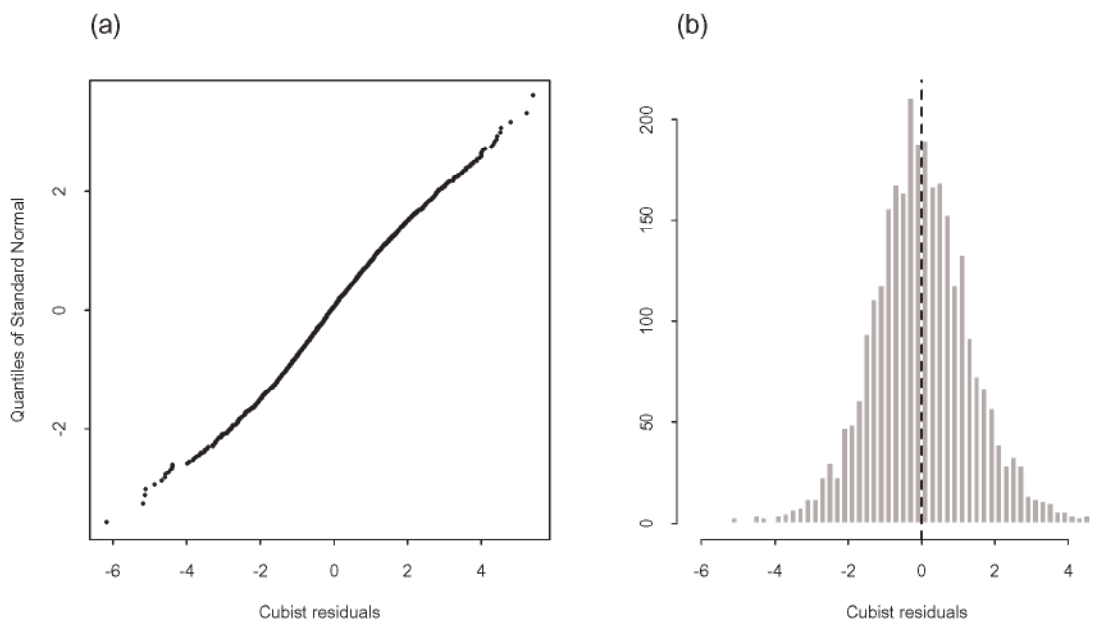


Figure 54: (a) Quantile plot and (b) histogram of Cubist residuals for sqrt(% clay) in layer 1 model (test data).

The model was then refitted using the same model options and variables on all 9750 observations. 32 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 0.99; relative error 0.65; correlation 0.72). The state and region-wise performance are summarized for this model in Tables 40 and 41.

Figure 53 can be decomposed into the 18 regions making up the ASRIS extent. This leads directly to Figure 55 and Figure 56 and enables some spatial assessment of performance.

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
NSW	2462	0.58	0.77	0.98	1.26
QLD	2824	0.73	0.62	1.00	1.29
SA	1213	0.55	0.77	0.96	1.26
TAS	463	0.56	0.74	1.12	1.45
VIC	499	0.62	0.72	0.97	1.29
WA	1586	0.46	0.87	0.89	1.16
CSIRO	703	0.69	0.69	0.88	1.08

Table 40: Performance of % clay in layer 1 model by state/CSIRO.

Region	N	rank correlation	relative error	average error	approximate RMSE
nterritory	20	0.50	1.03	1.76	2.00
carpentaria	8	0.95	0.46	0.67	0.85
qldnorth	56	0.68	0.68	1.01	1.22
qldcentral	661	0.61	0.74	1.05	1.34
qldsouth	506	0.75	0.58	0.94	1.22
moreton	157	0.52	0.78	1.10	1.38
burdekin	478	0.76	0.56	0.98	1.27
fitzroy	363	0.55	0.77	1.05	1.37
mdb	3094	0.73	0.62	0.95	1.23
nswnorth	493	0.51	0.86	1.08	1.38
nswsouth	692	0.32	0.94	0.87	1.13
viceast	134	0.51	0.87	0.79	1.02
vicwest	371	0.59	0.74	1.12	1.44
tasmania	463	0.56	0.74	1.12	1.45
sagulf	559	0.55	0.76	0.93	1.22
wplateau	119	0.36	0.94	0.93	1.17
wasouth	1540	0.45	0.88	0.89	1.16
indian	36	0.55	0.83	0.67	0.85

Table 41: Performance of % clay in layer 1 model by region.

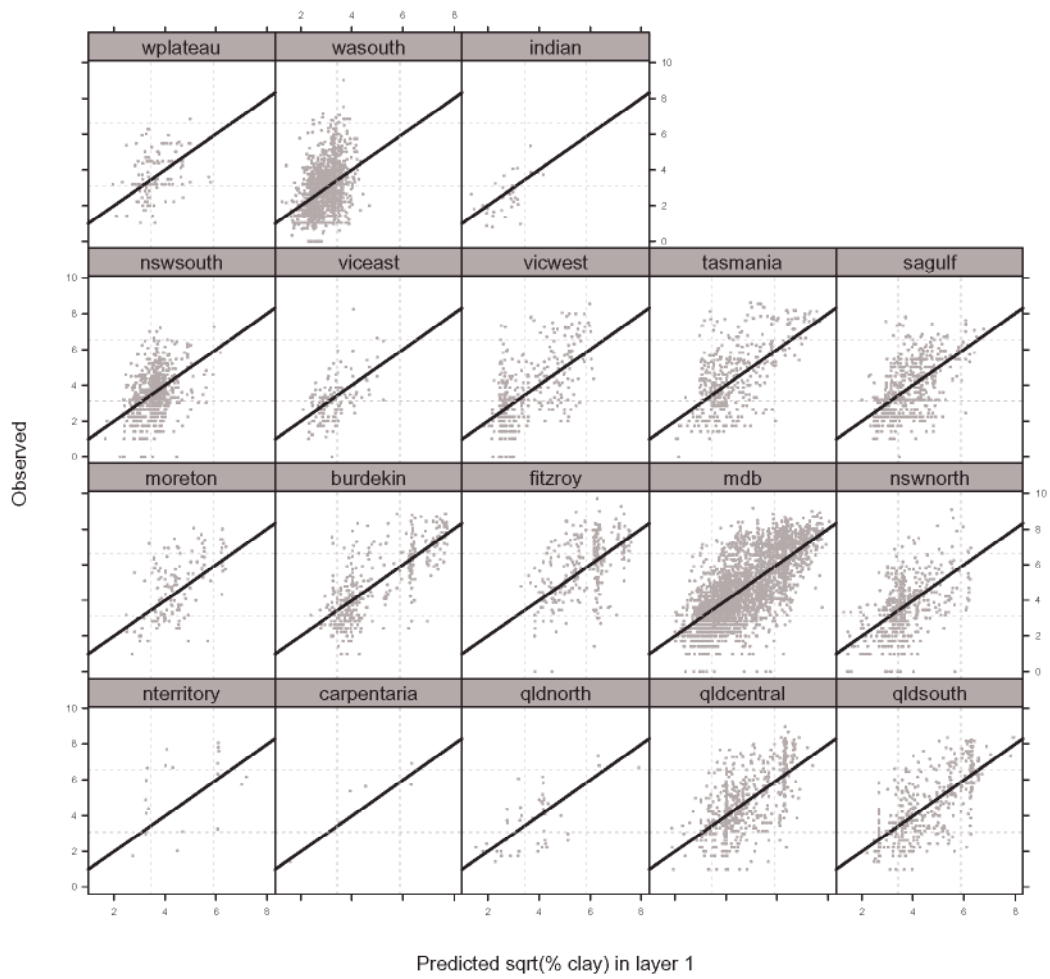


Figure 55: Observed versus predicted plots by region.

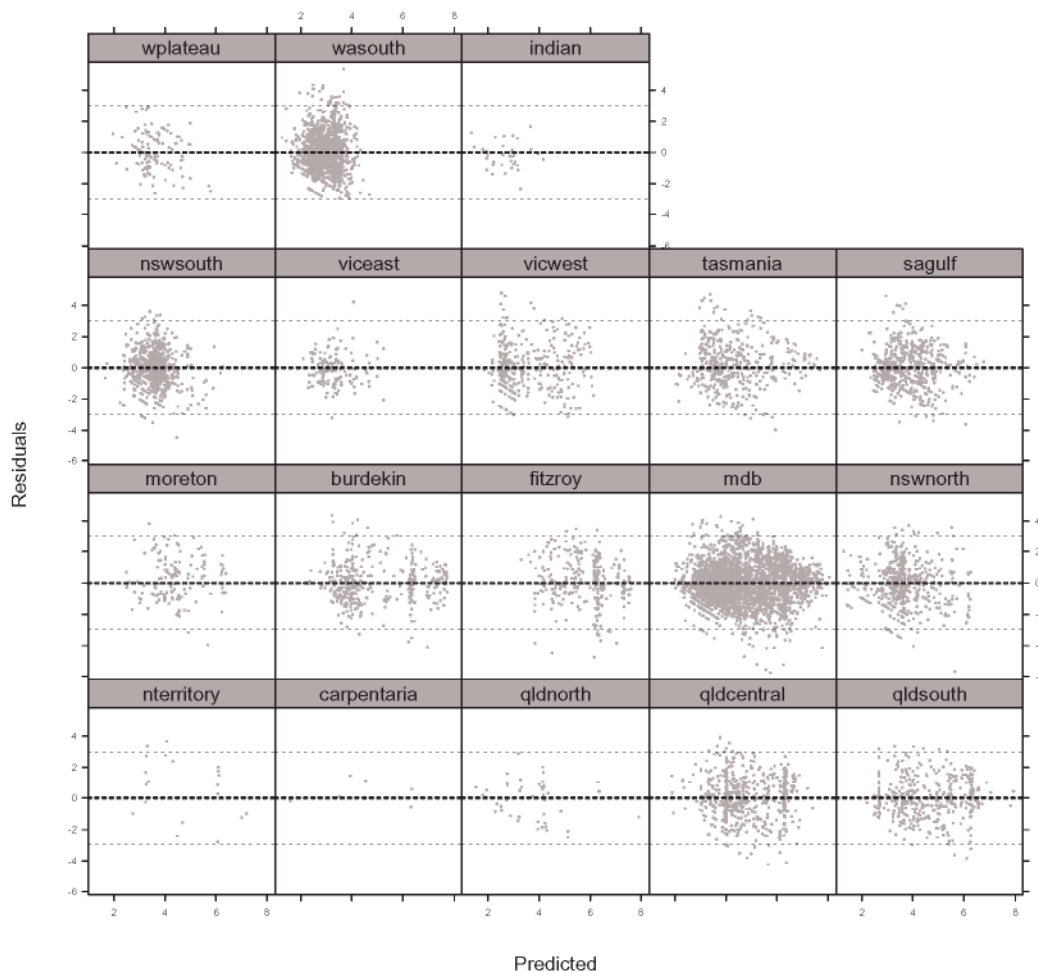


Figure 56: Residual plots by region.

The layer 1 square root % clay model appears to be relatively stronger in Queensland, eastern Victoria, Tasmania and the Murray-Darling basin. The performance in South Australia and northern New South Wales is fair, while the performance in Western Australia and southern New South Wales is not well supported. There is however clearly a large amount of unexplained variation in all regions.

The 32 rules from the final Cubist model were applied to ASRIS extent to generate a map of layer 1 % clay predictions. These predictions can be viewed at www.nlwra.gov.au/data/.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: relief, relative elevation, ridge distance, MSS band 4, highest period moisture index, annual mean radiation, annual mean precipitation and the maximum temperature of the warmest period. This certainty surface is available at www.nlwra.gov.au/data/.

8.2 Layer 2 clay

The distribution of layer 2 % clay by State/CSIRO is summarized in Table 42 and in the histograms in Figure 57. There is evidently a greater proportion of soils with low clay content in New South Wales and Western Australia. Tasmania and Victoria are both fairly uniformly distributed.

State	min	q10	q25	q50	q75	q90	max	N
NSW	0	5	16	30	46	59	94	2016
QLD	1	18	30	44	56	65	94	2376
SA	2	10	25	37	48	60	87	441
TAS	1	11	25	45	63	76	99	425
VIC	6	22	31	44	57	65	87	354
WA	0	8	18	32	43	53	80	1329
CSIRO	4	20	31	40	49	56	100	620

Table 42: Distribution by States/CSIRO.

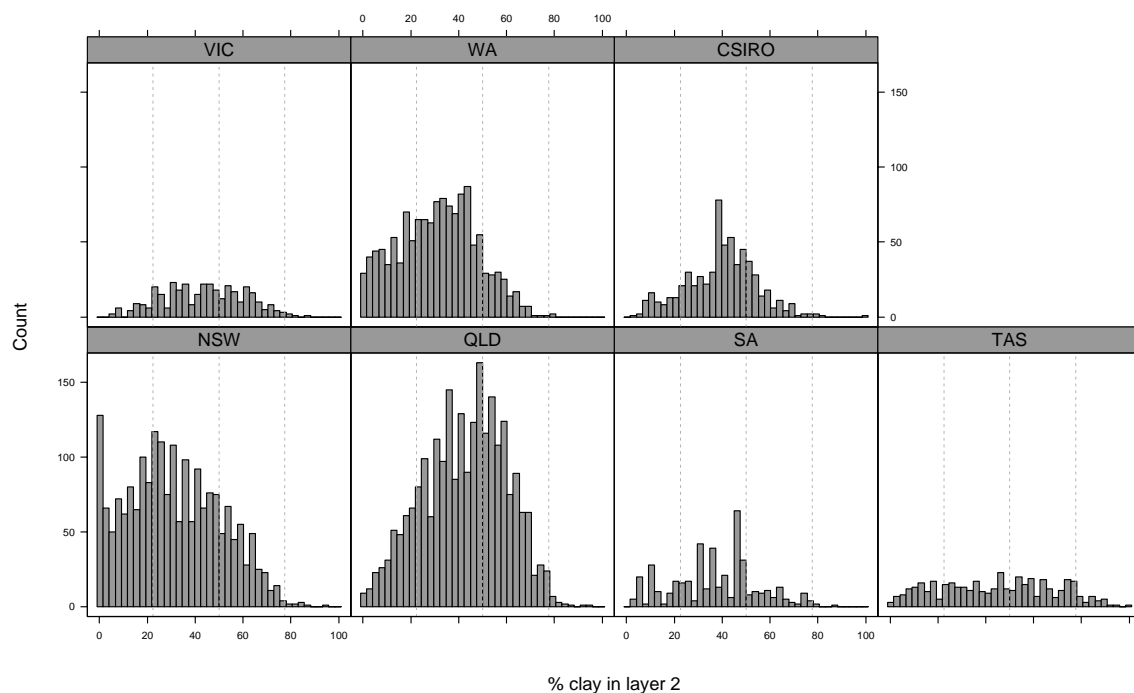


Figure 57: Histograms of layer 2 % clay by State.

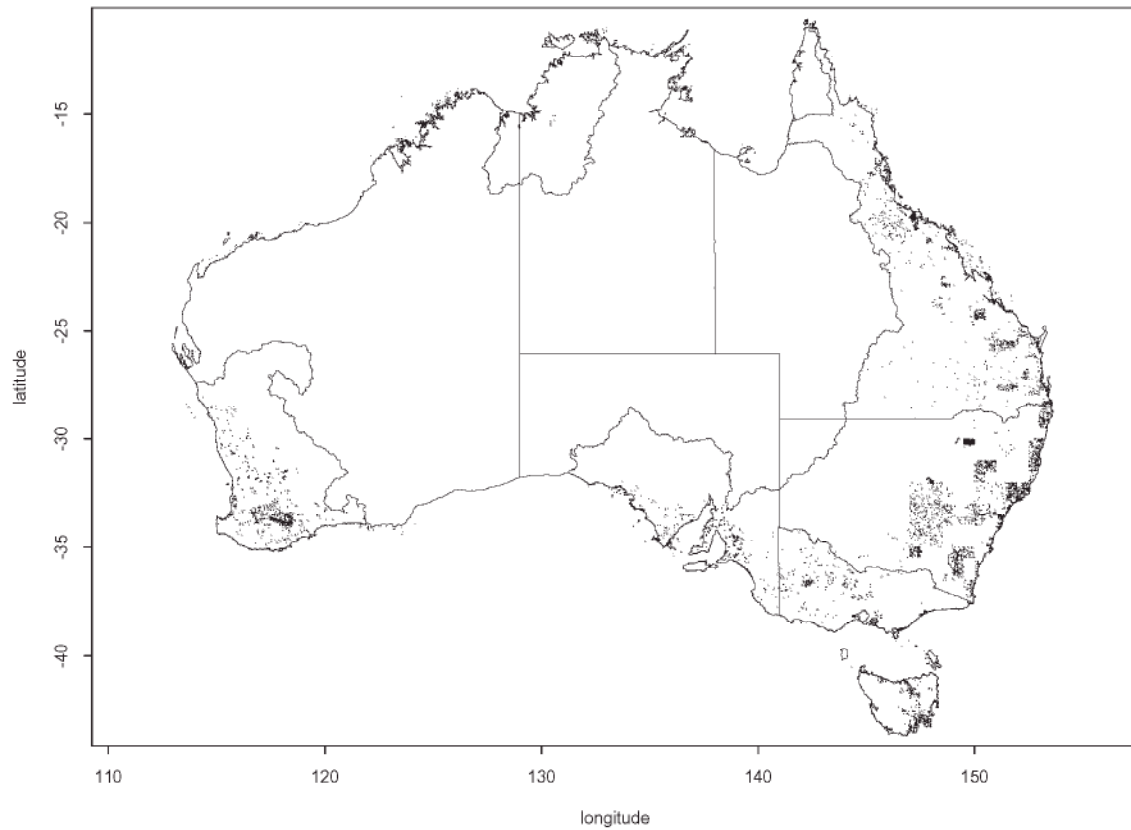


Figure 58: Locations of layer 2 % clay observations.

The locations of the 7050 layer 2 % clay observations used in the modelling are given in Figure 58.

A Cubist piecewise linear model was fitted to the square root of these data. 34 variables were used: 11 climatic, 3 MSS, 15 terrain, lithology, landuse and Atlas ASC and predicted % clay for layer 1 and layer 2 from a polygon model (via PPF derived look-up table). The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following diagnostics:

R^2	RMSE	average error	relative error	correlation
0.22	1.60	1.21	0.86	0.47

Table 43: sqrt(% clay) in layer 2 model diagnostics on test data set.

The overall performance on the test data is summarized graphically in Figure 59. The model is not as strong as layer 1 clay content. All performance indicators in Table 43 are notably worse than their layer 1 equivalents. There appears to be a fairly poor capacity to predict low clay contents.

The quantile plot and histogram of residuals are given in Figure 60.

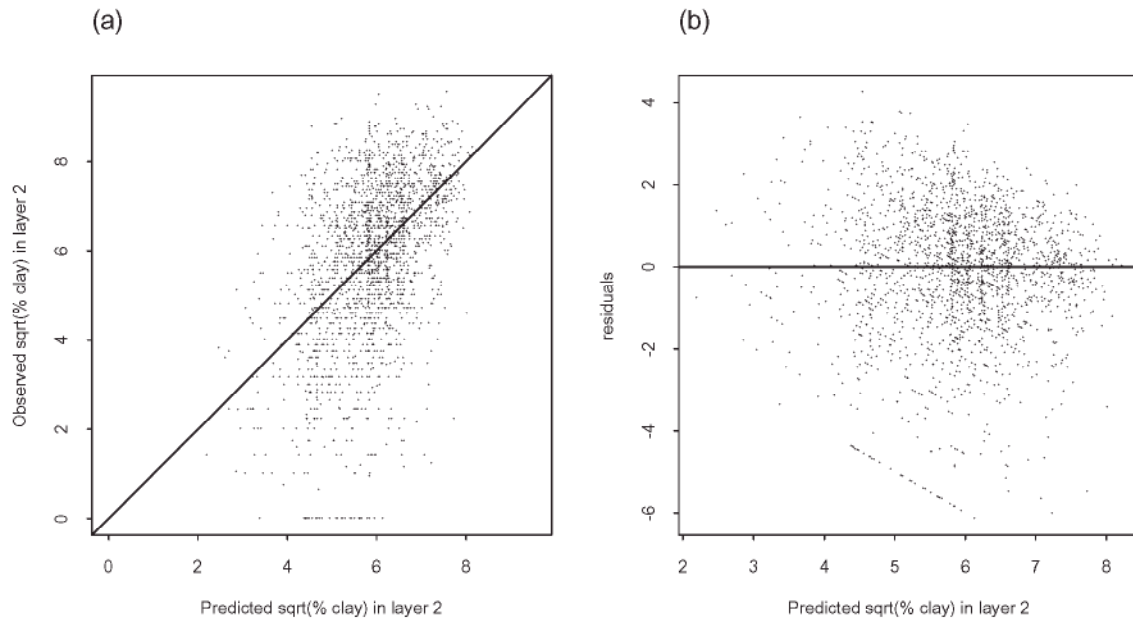


Figure 59: (a) Observed versus predicted and (b) residual plot for sqrt(% clay) in layer 2 model (test data only).

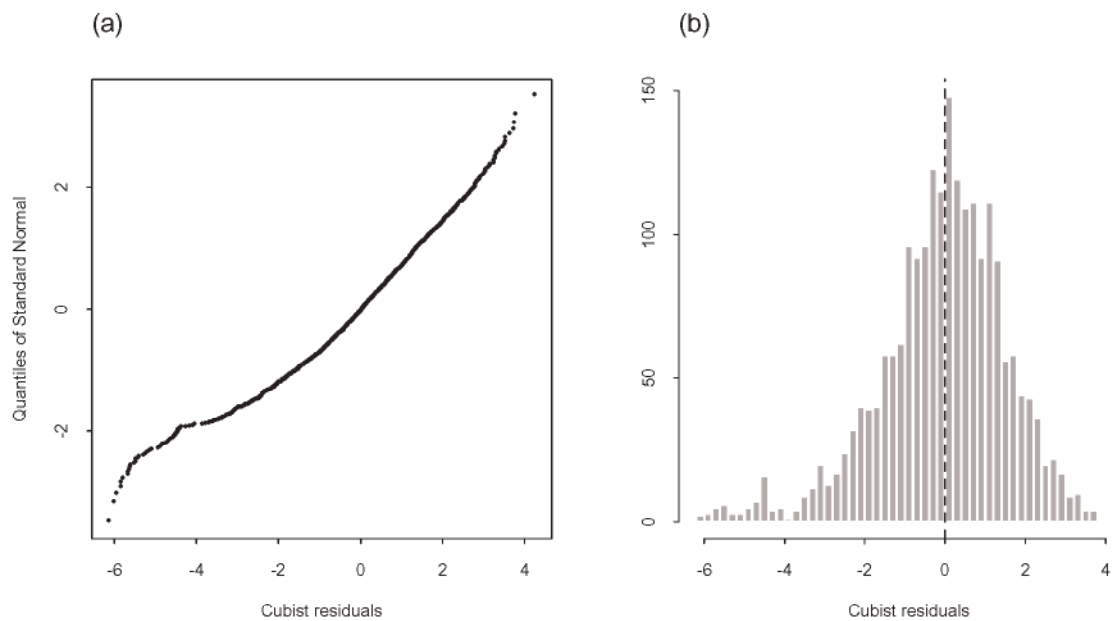


Figure 60: (a) Quantile plot and (b) histogram of Cubist residuals for sqrt(% clay) in layer 2 model.

The model was then refitted using the same model options and variables on all 7050 observations. 21 rules were used in the model. 10-fold cross-validation was performed on this model to judge performance (average error 1.14; relative error 0.81; correlation 0.54).

The State and region-wise performance are summarized in Tables 44 and 45.

Figure 59 can be decomposed into the 18 regions making up the ASRIS extent. This leads directly to Figure 61 and Figure 62 and enables some spatial assessment of performance.

State/CSIRO	N	rank correlation	relative error	average error	approximate RMSE
NSW	1931	0.45	0.87	1.43	1.86
QLD	2166	0.58	0.77	0.89	1.17
SA	424	0.39	0.89	1.13	1.48
TAS	350	0.54	0.81	1.26	1.61
VIC	342	0.44	0.87	0.96	1.24
WA	1240	0.45	0.84	1.13	1.47
CSIRO	597	0.40	0.85	0.81	1.10

Table 44: Performance of % clay in layer 2 model by state/CSIRO.

Region	N	rank correlation	relative error	average error	approximate RMSE
nterritory	18	-0.08	1.48	1.45	1.60
carpentaria	7	0.93	0.39	0.44	0.51
qldnorth	26	0.73	0.51	1.03	1.35
qldcentral	611	0.56	0.80	0.96	1.25
qldsouth	397	0.49	0.85	0.99	1.30
moreton	134	0.36	0.94	1.00	1.27
burdekin	339	0.61	0.78	0.86	1.10
fitzroy	230	0.42	0.90	0.86	1.18
mdb	2184	0.59	0.77	0.96	1.27
nswnorth	435	0.33	0.90	1.77	2.29
nswsouth	590	0.37	0.89	1.53	1.96
viceast	108	0.59	0.79	1.33	1.71
vicwest	88	0.16	0.97	1.04	1.35
tasmania	350	0.54	0.81	1.26	1.61
sagulf	262	0.33	0.94	1.15	1.50
wplateau	41	0.48	0.75	0.99	1.39
wasouth	1197	0.43	0.85	1.13	1.47
indian	33	0.48	0.95	1.14	1.48

Table 45: Performance of % clay in layer 2 model by region.

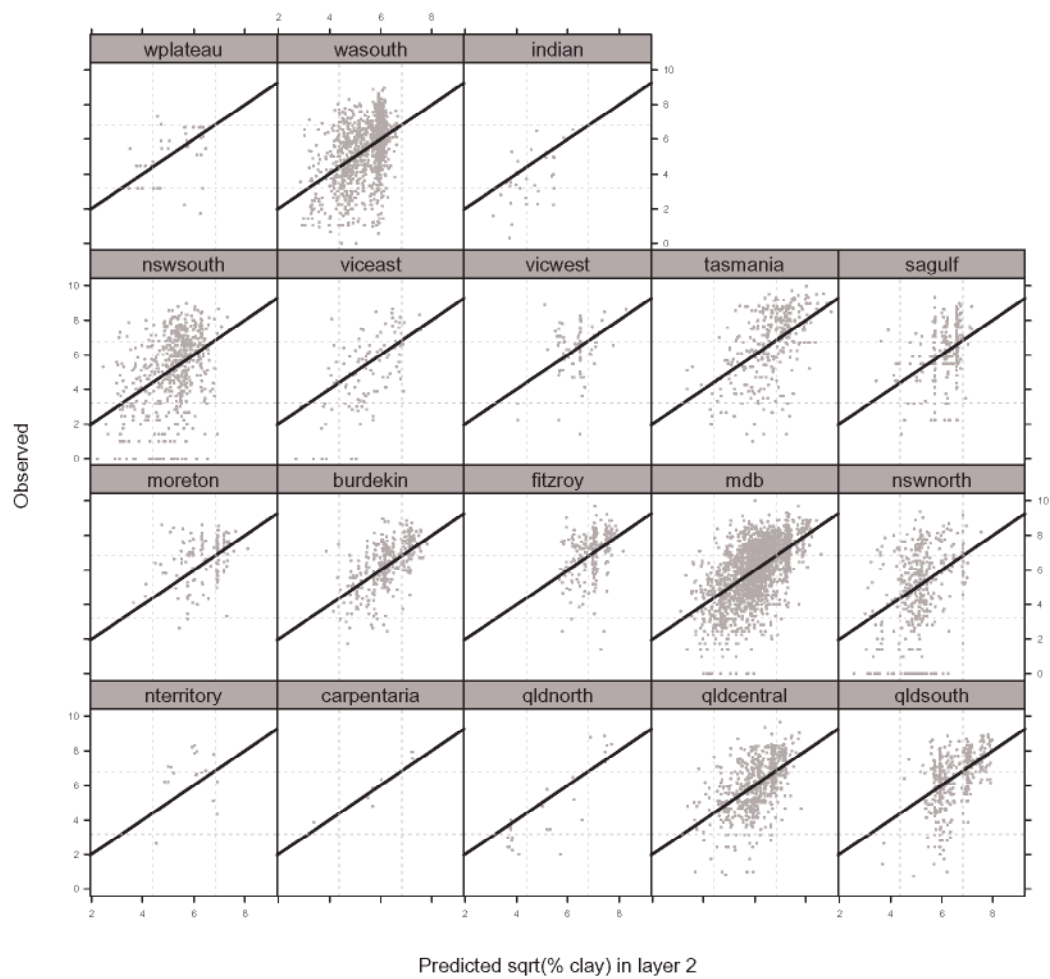


Figure 61: Observed versus predicted plots by region.

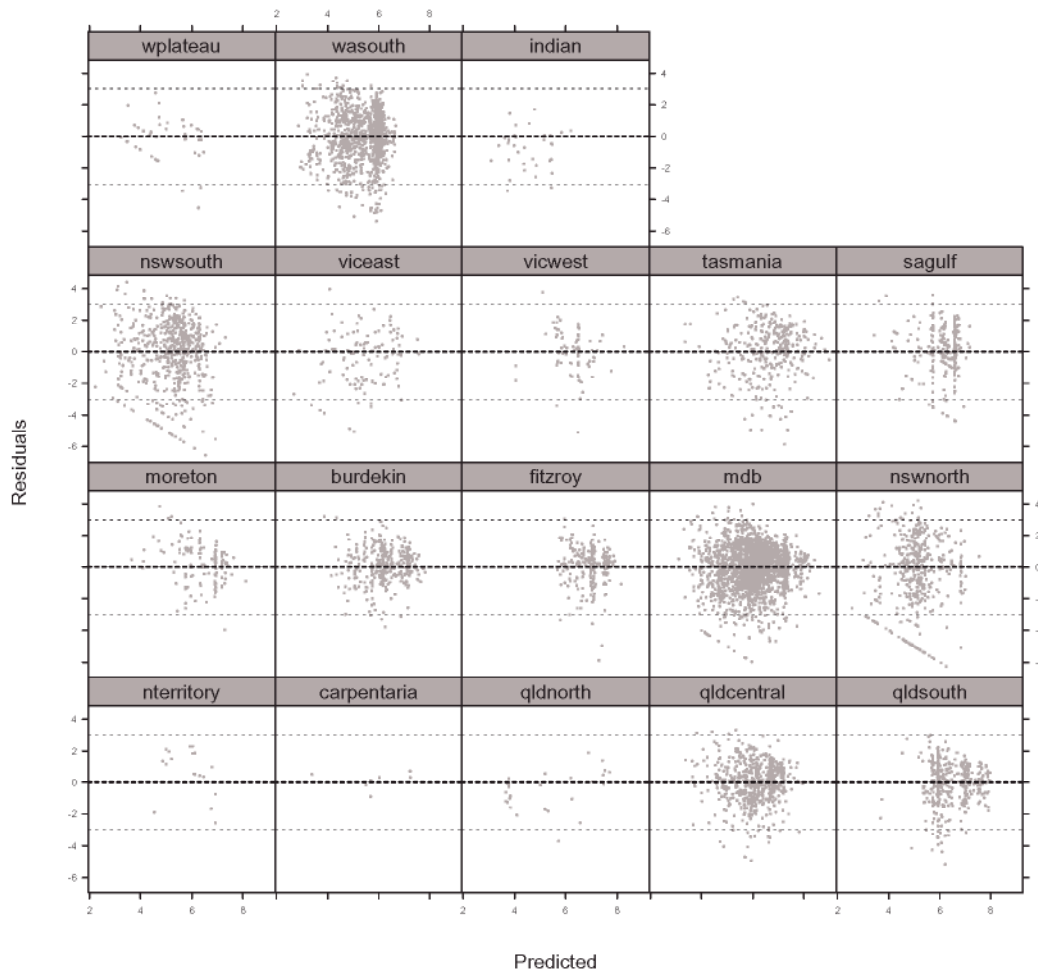


Figure 62: Residual plots by region.

There is weak support for this model. The prediction performance appears relatively stronger in south/central Queensland, the Murray-Darling basin, eastern Victoria and Tasmania. In all other regions the predictions are fairly poorly supported by the data.

The 21 rules from the final Cubist model were applied to ASRIS extent to generate a map of layer 2 % clay predictions. These predictions can be viewed at www.nlwra.gov.au/data/.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: relief, relative elevation, ridge distance, MSS band 4, highest period moisture index, annual mean radiation, annual mean precipitation and the maximum temperature of the warmest period. This certainty surface is available at www.nlwra.gov.au/data/.

9 TEXTURE

There were a large number of texture classes recorded in the ASRIS database, some 83 in layer 1 and 47 in layer 2. Many of these texture classes were however very poorly populated.

For the purposes of modelling it was decided to reduce the number of texture classes to 6, as described by McKenzie et al. (2000). The 6 levels used were: sands, sandy loams, loams, clay loams, light clays and clays.

9.1 Layer 1 texture

The counts across the 6 texture classes by State/CSIRO are given in Table 46. The features that stand out the most in this table are the great majority of sands in Western Australia and to a lesser extent South Australia and the larger proportion of clay loams and light clay textures in Queensland and Tasmania.

Class	NSW	QLD	SA	TAS	VIC	WA	CSIRO
sands	1961	3817	624	382	375	22347	129
sandy loams	3257	7640	453	934	762	7262	413
loams	5752	5148	435	550	535	5326	476
clay loams	3348	10213	157	1213	247	800	404
light clays	646	9378	125	376	249	862	315
clays	1912	6425	147	184	195	225	235
Total	16876	42621	1941	3639	2363	36822	1972

Table 46: Texture class counts by State/CSIRO.

The location of the 99316 layer 1 texture observations that were used in the modelling are given in Figure 63.

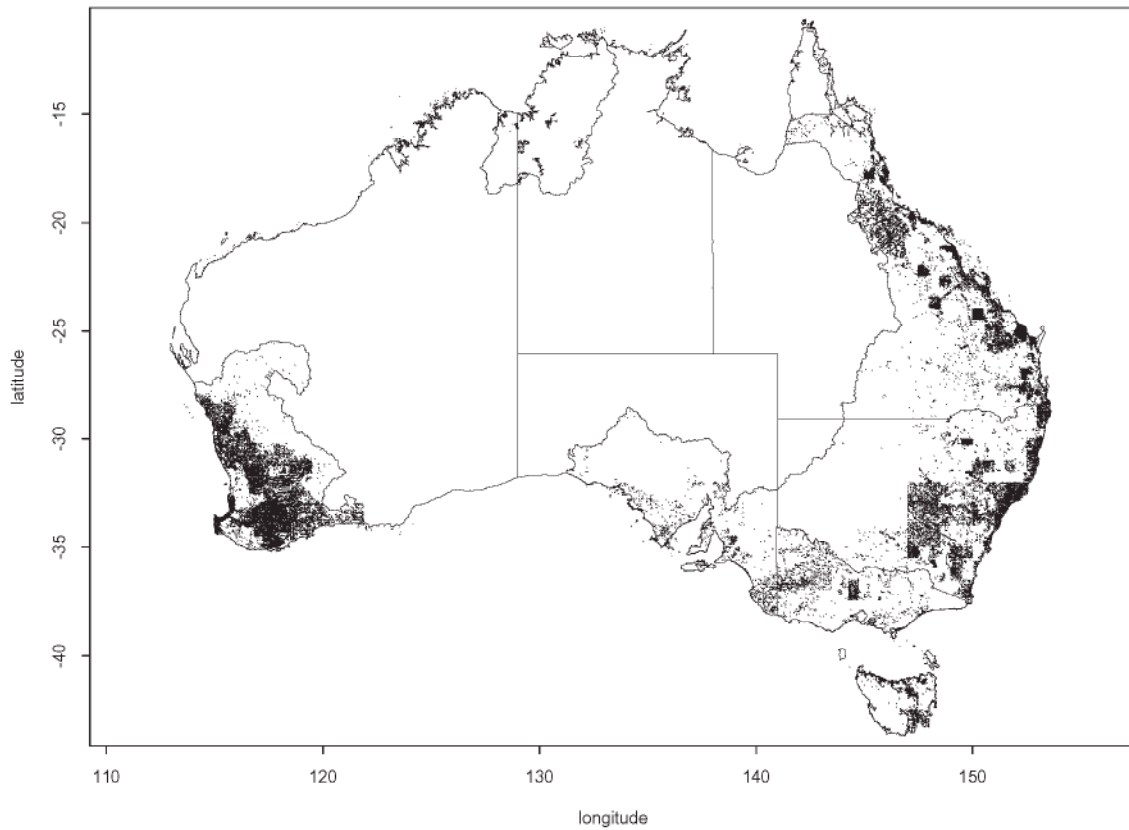


Figure 63: Locations of layer 1 texture observations.

C5.0 classification models were investigated for these data. Some difficulty was found in predicting the 5th class in the 6 class texture classification model. It was subsequently decided to pool “clay loams” and the “light clays” to create a new 5 class texture model. A *C5.0* model was fitted to these data. 37 variables were used: 11 climatic, 3 MSS, 16 terrain, lithology, landuse, ASC and the polygon predictions (via PPF look-up tables; see Carlile et al., 2001) for percent clay and sand in layers 1 and 2. The initial model was constructed with a 70:30 train:test split. The overall classification rate, which is the observed probability of correct classification (i.e. $1 - \text{error rate}$), for the 29796 observations in the test data set is 0.535. The model yields the following confusion matrix for the test data, where *c/rate* denotes the class-specific classification rate.

Observed		Predicted					Total	c/rate
		I	II	III	IV	V		
I	sands	6802	432	256	558	32	8080	0.84
II	sandy loams	2377	1012	627	1677	84	5777	0.17
III	loams	1728	496	1230	1632	101	5187	0.24
IV	clay loam, light clay	595	517	578	5711	646	8047	0.71
V	clay	103	76	134	1203	1189	2705	0.44

Table 47: Confusion matrix for layer 1 texture (5 class texture model).

The model does well at identifying the sands and (clay loam, light clays) but does fairly poorly at distinguishing the sandy loams and the loams from adjacent categories. While these classification rates are fairly low, some encouragement should be drawn from the fact that sands – loams (classes I, II, III) are unlikely to be classified as clays and vice versa.

The model was refitted using the same model options and variables for all layer 1 texture observations. 83 rules were used in the model.

Tables 48 and 49 present the prediction performance across the States and regions respectively. The classification rate is given both overall and for each texture category respectively.

There is a relatively weak ability to detect some large texture classes represented in some of the States/regions. For example, for Western Australia anything other than category I (sands) is very poorly predicted, despite some large counts for categories II and III.

State	overall		class I		class II		class III		class IV		class V	
	N	c/r	N	c/r	N	c/r	N	c/r	N	c/r	N	c/r
NSW	16876	0.43	1961	0.45	3257	0.18	5752	0.65	3994	0.32	1912	0.41
QLD	41854	0.54	3622	0.38	7423	0.26	5075	0.01	19404	0.85	6330	0.44
SA	1762	0.38	602	0.72	410	0.34	363	0.10	252	0.10	135	0.21
TAS	3499	0.54	318	0.10	898	0.57	518	0.12	1584	0.81	181	0.00
VIC	2363	0.36	375	0.53	762	0.22	535	0.30	496	0.40	195	0.63
WA	35166	0.61	21352	0.99	6958	0.01	5197	0.05	1453	0.02	206	0.00
CSIRO	1837	0.44	98	0.33	390	0.14	440	0.43	689	0.61	220	0.53

Table 48: Classification rate by State/CSIRO for layer 1 texture (5 class texture model).

Region	overall		class I		class II		class III		class IV		class V	
	N	c/r	N	c/r	N	c/r	N	c/r	N	c/r	N	c/r
nterritory	659	0.32	50	0.70	83	0.00	106	0.00	275	0.59	145	0.12
carpentaria	222	0.45	56	0.96	48	0.00	30	0.03	87	0.52	1	0.00
qldnorth	596	0.57	222	0.96	139	0.00	32	0.00	187	0.66	16	0.00
qldcentral	9342	0.51	770	0.50	1712	0.02	1394	0.00	4557	0.91	909	0.22
qldsouth	11350	0.54	1411	0.34	3048	0.56	1335	0.00	5032	0.79	524	0.00
moreton	5773	0.57	246	0.07	766	0.09	713	0.00	3182	0.88	866	0.48
burdekin	7473	0.54	609	0.33	922	0.05	1044	0.04	3051	0.86	1847	0.59
fitzroy	6287	0.58	249	0.04	735	0.06	479	0.00	3103	0.90	1721	0.45
mdb	11663	0.43	905	0.30	2613	0.26	3198	0.53	2847	0.41	2100	0.59
nswnorth	2371	0.44	313	0.59	210	0.18	809	0.64	736	0.41	303	0.00
nswsouth	6406	0.43	1034	0.55	1266	0.07	2413	0.72	1409	0.19	284	0.20
viceast	886	0.36	176	0.32	245	0.15	260	0.55	190	0.41	15	0.00
vicwest	740	0.44	275	0.73	176	0.12	85	0.39	103	0.29	101	0.41
tasmania	3497	0.54	317	0.10	898	0.57	518	0.12	1583	0.81	181	0.00
sagulf	827	0.36	262	0.63	225	0.51	196	0.06	117	0.04	27	0.04
wplateau	291	0.31	92	0.83	66	0.00	89	0.15	25	0.04	19	0.00
wasouth	32576	0.62	19911	0.99	6526	0.01	4766	0.04	1258	0.01	115	0.00
indian	2398	0.62	1430	0.97	420	0.01	413	0.17	130	0.15	5	0.20

Table 49: Classification rate by region for layer 1 texture (5 class texture model).

The conclusions that can be made from Tables 48 and 49 are that:

- Class I (sands) are best identified in Western Australia, South Australia, western Victoria, New South Wales and far north Queensland. The observations of texture class sand are not as well identified in south/central Queensland and Tasmania.
- Class II (sandy loams) are well predicted in Tasmania, South Australia and some of southern Queensland. They are particularly poorly identified in Western Australia.
- Class III (loams) are classified correctly more often in New South Wales and Victoria but often incorrectly in Western Australia and Queensland.
- Class IV (clay loams/light clays) are well identified in Queensland, Northern Territory and Tasmania but poorly in Western Australia and southern NSW.
- There are not a lot of class V (clay) records in layer 1. Those in the Murray-Darling basin and south/central Queensland are generally well predicted.

The 83 rules from the final *C5.0* model were applied to ASRIS extent to generate a map of layer 1 soil texture predictions. These predictions can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: relief, relative elevation, ridge distance, MSS band 4, highest period moisture index, annual mean radiation, annual mean precipitation and the maximum temperature of the warmest period. This certainty surface is available at www.nlwra.gov.au/data.

9.2 Layer 2 texture

The counts across the 6 texture classes by State/CSIRO are given in Table 50. With the exception of Western Australia, clays are the most dominant class. Queensland clearly provides a large proportion of the counts.

Class	NSW	QLD	SA	TAS	VIC	WA	CSIRO
sands	542	735	72	115	29	1930	21
sandy loams	306	594	64	90	44	1376	39
loams	1702	1195	157	161	75	3335	76
clay loams	1060	1358	27	101	52	828	144
light clays	44	12879	301	722	416	5259	614
clays	9129	24099	654	2073	1194	3166	669
Total	12783	40860	1275	3262	1810	15894	1563

Table 50: Texture class counts by State/CSIRO.

The locations of the 73163 layer 2 texture observations used in the modelling are given in Figure 64.

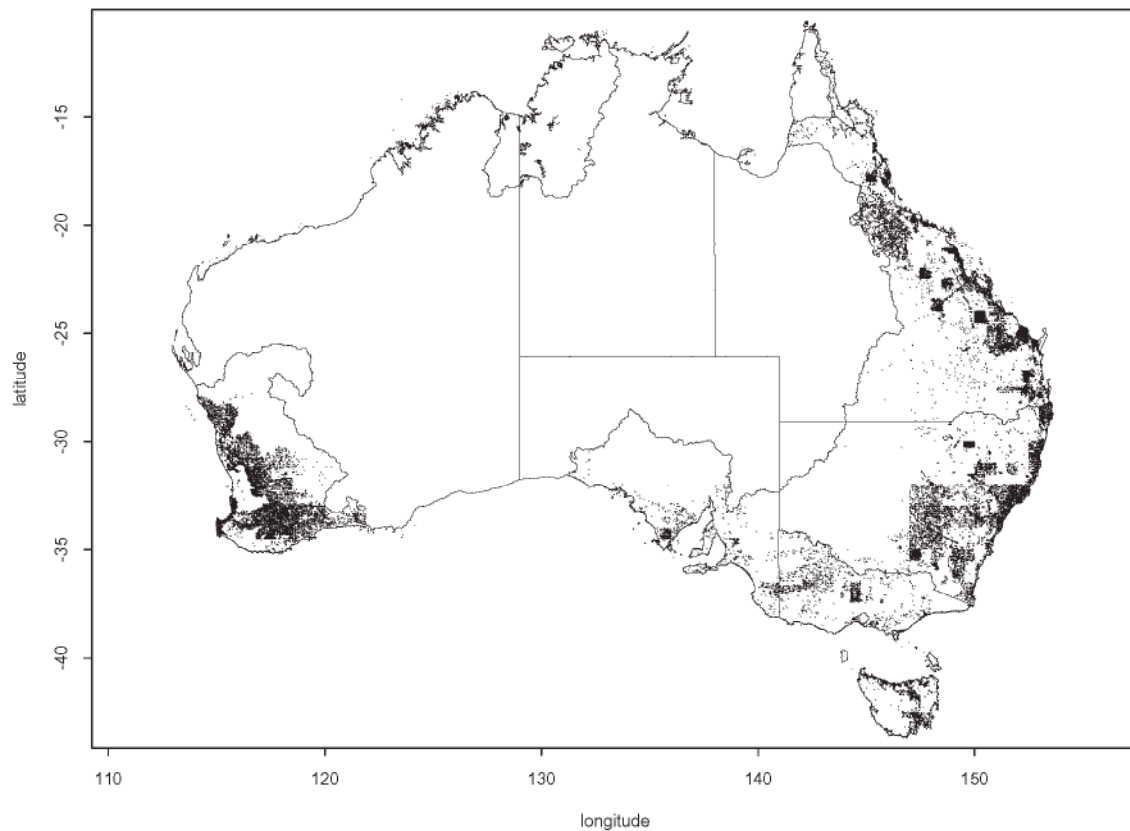


Figure 64: Locations of layer 2 texture observations.

As for layer 1 texture, it was difficult to construct a sensible classification model for the 6 texture classes in Table 50. It was decided to pool the first three classes, the fourth and fifth and leave the sixth class as it is to create a 3 category texture classification.

A *C5.0* model was fitted to these data. 37 variables were used: 11 climatic, 3 MSS, 16 terrain, lithology, landuse, ASC and the polygon predictions (via PPF look-up tables) for percent clay and sand in layers 1 and 2. The initial model was constructed with a 70:30 train:test split. The overall error rate for the 21949 observations in the test data set is 0.329. The model yields the following confusion matrix for the test data.

Observed		Predicted			total	c/rate
		a	b	c		
a	sands, sandy loams, loams	1916	706	847	3469	0.552
b	clay loams, light clays	722	3123	2845	6690	0.467
c	clays	517	1574	9699	11790	0.823

Table 51: Confusion matrix for layer 2 texture (3 class texture model).

The model does well at identifying clay. The most difficulty is in distinguishing the texture category with clay loams and light clays from that of the clays.

The model was refitted using the same model options and variables for all layer 2 texture observations. 59 rules were used in the model. Tables 52 and 53 give the performance across the State and ASRIS region for this model.

State	overall		class a		class b		class c	
	N	c/r	N	c/r	N	c/r	N	c/r
NSW	12783	0.74	2550	0.34	1104	0.05	9129	0.93
QLD	40221	0.67	2371	0.18	13929	0.47	23921	0.84
SA	1179	0.56	212	0.17	317	0.33	650	0.79
TAS	3155	0.68	320	0.08	800	0.40	2035	0.88
VIC	1810	0.64	148	0.03	468	0.10	1194	0.93
WA	15130	0.58	6342	0.78	5675	0.51	3113	0.31
CSIRO	1522	0.56	128	0.04	737	0.33	657	0.92

Table 52: Classification rate by State/CSIRO for layer 2 texture (3 class texture model).

Region	overall		class a		class b		class c	
	N	c/r	N	c/r	N	c/r	N	c/r
nterritory	512	0.57	55	0.00	193	0.28	264	0.90
carpentaria	191	0.32	56	0.05	76	0.39	59	0.47
qldnorth	554	0.44	167	0.19	271	0.61	116	0.41
qldcentral	10055	0.74	666	0.19	2770	0.41	6619	0.93
qldsouth	10663	0.63	736	0.26	4545	0.61	5382	0.69
moreton	5267	0.61	200	0.16	2173	0.48	2894	0.74
burdekin	6934	0.71	407	0.12	1933	0.42	4594	0.89
fitzroy	5933	0.66	128	0.03	2028	0.25	3777	0.90
mdb	9017	0.74	921	0.08	1530	0.18	6566	0.96
nswnorth	1821	0.70	320	0.39	279	0.04	1222	0.94
nswsouth	4888	0.71	1363	0.47	368	0.03	3157	0.89
viceast	632	0.46	220	0.23	132	0.17	280	0.78
vicwest	357	0.52	37	0.08	146	0.32	174	0.77
tasmania	3154	0.68	319	0.08	800	0.40	2035	0.88
sagulf	604	0.68	84	0.17	112	0.29	408	0.90
wplateau	227	0.50	56	0.25	21	0.43	150	0.60
wasouth	13766	0.57	5448	0.75	5446	0.53	2872	0.29
indian	1225	0.73	888	0.95	207	0.20	130	0.06

Table 53: Classification rate by region for layer 2 texture (3 class texture model).

The classification model does particularly well at identifying the layer 2 texture observations of class clay throughout the extent, with the exception of Western Australia and far north Queensland.

Class “a” (sands, sandy loams, loams) have a reasonable classification rate in Western Australia and southern New South Wales, but a fairly poor rate in the Murray-Darling basin and south/central Queensland.

Class “b” (clay loams/light clays) are generally well-identified in Western Australia and Queensland but not so well recognised in the New South Wales.

The 59 rules from the final *C5.0* model were applied to ASRIS extent to generate a map of layer 2 soil texture predictions. These predictions can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: relief, relative elevation, ridge distance, MSS band 4, highest period moisture index, annual mean radiation, annual mean precipitation and the maximum temperature of the warmest period. This certainty surface is available at www.nlwra.gov.au/data.

10 THICKNESS

Layer 1 and layer 2 thickness were reported in both centimetres and metres. While units used were largely consistent within the agency, some agencies clearly reported both centimetres and metres. It was necessary to convert all thickness measurements to metres prior to analysis.

Thicknesses reported as 0 were omitted if measurements for other properties in the same layer existed. For example, if a layer 2 thickness was recorded as 0 but a pH assessment given, it was assumed that the thickness was not actually 0 but rather missing.

There was some censoring in this data, especially in the layer 2 thickness, where depth limitations imposed by the equipment or a decision on a sufficient depth resulted in an incomplete assessment of thickness. It was however not known which observations were censored in the database. Given there were some very deep thicknesses recorded it was decided to clip the thicknesses back to 2.5 metres in layer 1 and 5.0 metres in layer 2. These values are used in Tables 54 and 59.

10.1 *Layer 1 thickness*

State	min	q10	q25	q50	q75	q90	max	N
NSW	0.00	0.10	0.15	0.25	0.35	0.55	2.5	17276
QLD	0.01	0.06	0.10	0.20	0.30	0.50	2.5	43659
SA	0.00	0.10	0.15	0.25	0.39	0.57	2.5	20169
TAS	0.00	0.12	0.17	0.23	0.33	0.50	2.5	3679
VIC	0.00	0.10	0.10	0.20	0.35	0.53	2.0	2213
WA	0.00	0.10	0.15	0.30	0.50	0.80	2.5	23715
CSIRO	0.00	0.05	0.10	0.20	0.30	0.42	2.5	1748

Table 54: Distribution by States/CSIRO.

Figure 65 presents the histograms of the layer 1 thickness. Recorded thicknesses greater than 1.5 metres are truncated to 1.5 metres for purposes of illustration. Some digit preference is evident with a tendency of the thicknesses to cluster around certain values. The locations of these layer 1 thickness observations is given in Figure 66.

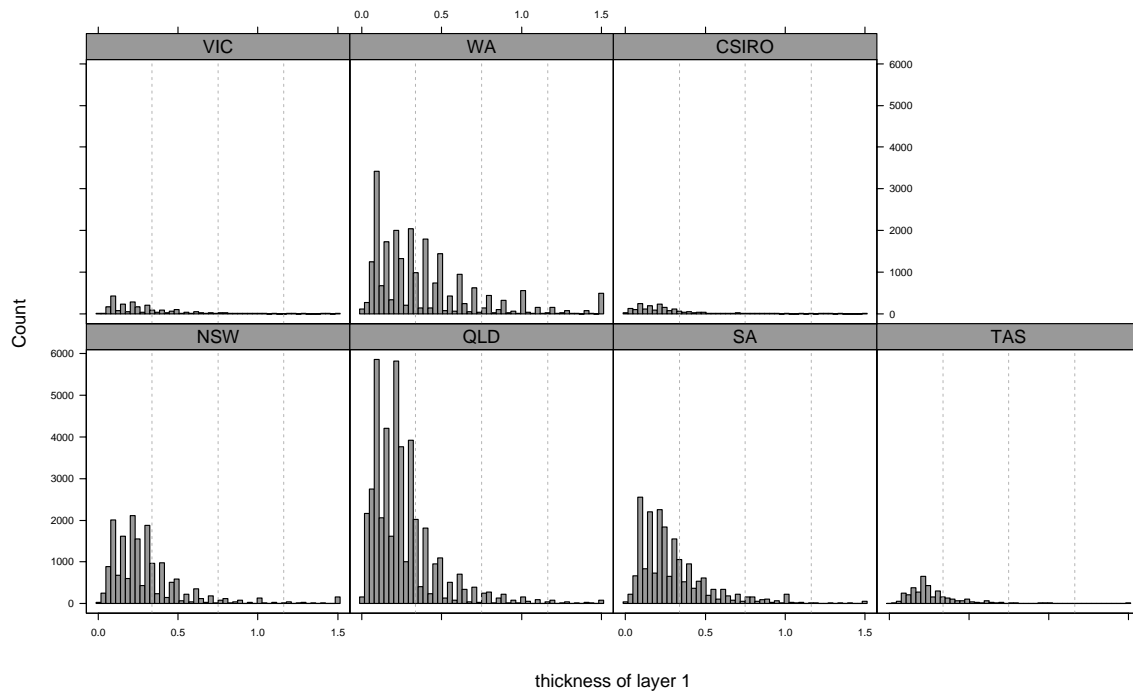


Figure 65: Histograms of layer 1 thickness by State.

Cubist piecewise linear models were investigated for the continuous thickness of layer 1. Unfortunately, those models constructed did not show good predictive ability. It was then decided to consider layer 1 thickness as a two category variable with levels thin/thick, where thin was defined by having a layer 1 thickness < 0.25 metres. The choice of 0.25 metres was made as that was close to the median layer 1 thickness which ensured that approximately half of the thickness measurements were deemed thin.

Thickness	NSW	QLD	SA	TAS	VIC	WA	CSIRO
thin ($< 0.25m$)	8478	25058	9979	1996	1231	9934	1112
thick ($\geq 0.25m$)	8798	18601	10190	1683	982	13781	636
total	17276	43659	20169	3679	2213	23715	1748

Table 55: Layer 1 thin/thick by State.

A *C5.0* classification model was fitted to these data. 34 variables were used: 11 climate, 3 MSS, 16 terrain, lithology, landuse, ASC and the predicted layer 1 thickness as derived from a polygon method (via PPF derived look-up table). The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the confusion matrix in Table 56.

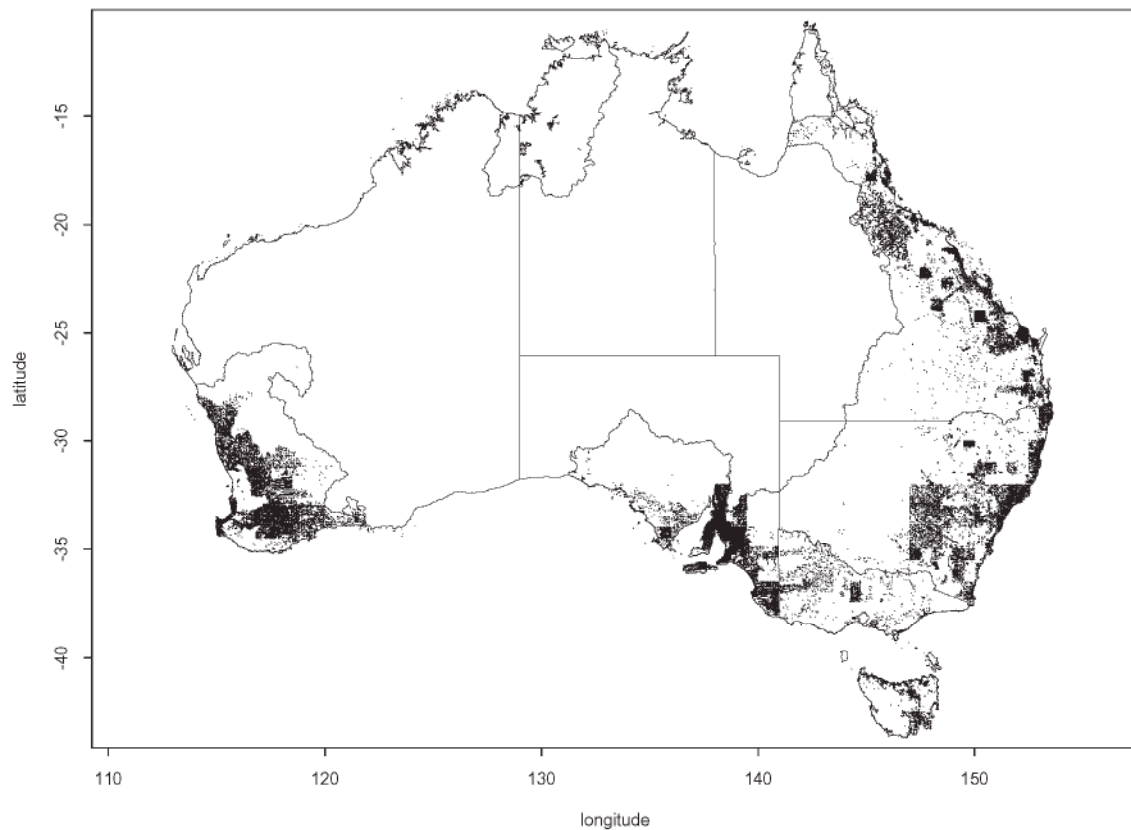


Figure 66: Locations of layer 1 thickness observations.

		<i>Predicted</i>	
		thin	thick
<i>Observed</i>	thin	11293	5137
	thick	5494	9919

Table 56: Observed versus predicted for two category thickness in layer 1.

There are 31843 observations in the test data. The overall classification rate, which is the observed probability of correct classification (i.e. $1 - \text{error rate}$), is 0.666. The population specific classification rates are 0.687 for thinner and 0.644 for the thicker layer 1 thickness measurements.

Boosting was investigated but was found to improve the predictive power negligibly, while considerably increasing the computational demands in the prediction step. It was not used in the final model.

The model was then refitted using the same model options and variables on all layer 1 observations. 53 rules were used in the final model.

Tables 57 and 58 give the classification rates, both overall and for classes thin and thick, for this final model across all States and regions.

State	overall		thin		thick	
	N	c/rate	N	c/rate	N	c/rate
NSW	17276	0.60	8478	0.52	8798	0.68
QLD	42982	0.72	24705	0.84	18277	0.56
SA	19455	0.69	9533	0.71	9922	0.67
TAS	3535	0.62	1939	0.64	1596	0.59
VIC	2213	0.62	1231	0.68	982	0.54
WA	22774	0.64	9488	0.45	13286	0.78
CSIRO	1704	0.67	1075	0.72	629	0.59

Table 57: Classification rate by State for layer 1 thickness (binary thickness model).

Region	overall		thin		thick	
	N	c/rate	N	c/rate	N	c/rate
nterritory	611	0.75	472	0.78	139	0.64
carpentaria	218	0.58	116	0.23	102	0.98
qldnorth	570	0.66	235	0.31	335	0.91
qldcentral	9668	0.71	5438	0.78	4230	0.61
qldsouth	11627	0.73	4205	0.61	7422	0.80
moreton	6036	0.65	3859	0.89	2177	0.23
burdekin	7872	0.73	5525	0.94	2347	0.22
fitzroy	6291	0.79	4998	0.98	1293	0.08
mdb	15751	0.65	8581	0.71	7170	0.59
nswnorth	2509	0.58	1118	0.39	1391	0.73
nswsouth	6614	0.58	2887	0.33	3727	0.77
viceast	1016	0.62	383	0.36	633	0.78
vicwest	2489	0.70	1028	0.51	1461	0.84
tasmania	3533	0.62	1939	0.64	1594	0.59
sagulf	11877	0.68	5797	0.73	6080	0.64
wplateau	671	0.66	466	0.86	204	0.19
wasouth	20816	0.65	8572	0.45	12244	0.79
indian	1771	0.61	830	0.56	941	0.66

Table 58: Classification rate by region for layer 1 thickness (binary thickness model).

The classification rates for thick and thin at times appear quite unbalanced. For example, in Fitzroy 98% of the 4998 thin soils are classified correctly but only 8% of the 1293 thicker soils.

Tables 57 and 58 suggest that in northern Queensland the model performs markedly better predicting the thick layer 1 soils, while in southern/central Queensland the model is generally better at predicting the thin soils. The Murray-Darling Basin (MDB) is marginally better at identifying thin soils. For non-MDB New South Wales and Victoria thick, layer 1 soils are notably better recognised than thin soils. Tasmania and South Australia are fairly consistent for both thin and thick soils. Western Australia is stronger at identifying thick soils.

The 53 rules from the final *C5.0* model were applied to ASRIS extent to generate a map of layer 1 thickness class predictions. These predictions can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: relief, relative elevation, ridge distance, MSS band 4, highest period moisture index, annual mean radiation, annual mean precipitation and the maximum temperature of the warmest period. This certainty surface is available at www.nlwra.gov.au/data

10.2 Layer 2 thickness

State	min	q10	q25	q50	q75	q90	max	N
NSW	0.00	0.20	0.32	0.52	0.80	1.25	5.00	15326
QLD	0.00	0.27	0.45	0.80	1.18	1.50	5.00	43058
SA	0.00	0.13	0.25	0.45	0.66	0.85	4.98	17934
TAS	0.00	0.20	0.34	0.49	0.64	0.90	4.52	3471
VIC	0.00	0.22	0.35	0.59	0.90	1.25	5.00	1790
WA	0.00	0.13	0.20	0.40	0.70	1.08	5.00	18749
CSIRO	0.03	0.25	0.40	0.70	1.12	1.46	4.25	1611

Table 59: Distribution by State/CSIRO.

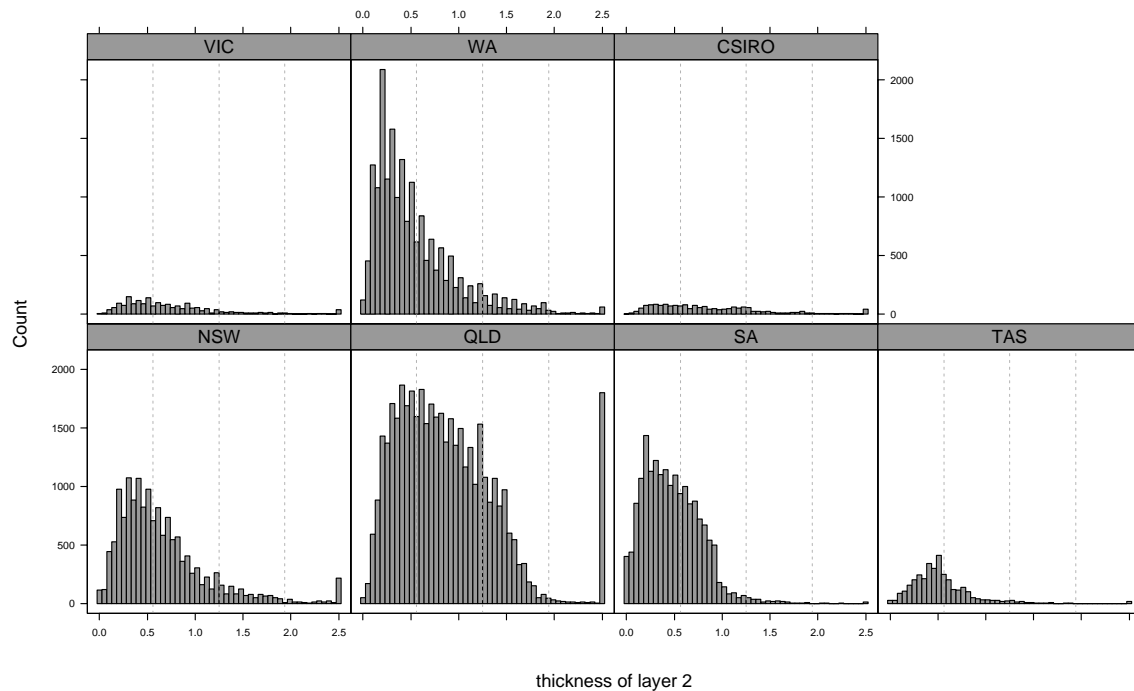


Figure 67: Histograms of layer 2 thickness by State.

Figure 67 presents the histograms of the layer 2 thickness. Recorded thicknesses greater than 2.5 metres were truncated to 2.5 metres for presentation. This is most notable in Queensland where there appear to be a much larger proportion of deep soils measured.

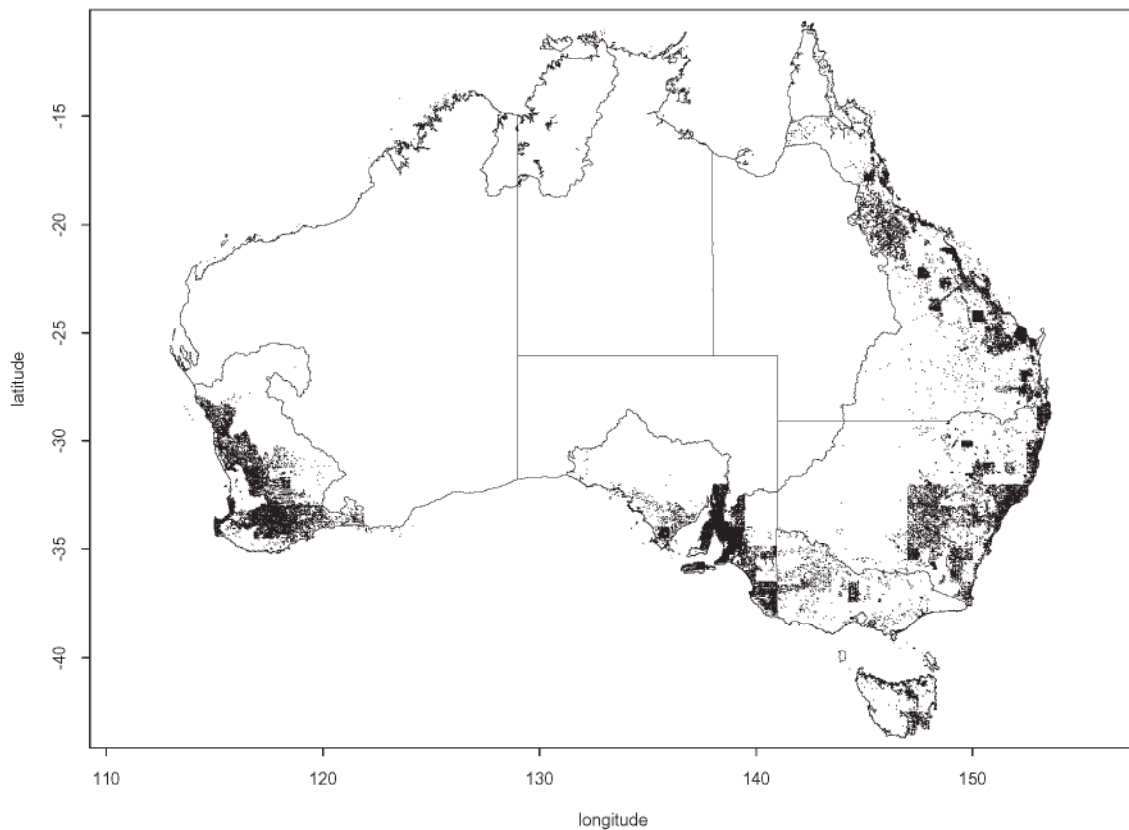


Figure 68: Locations of layer 2 thickness observations

Cubist piecewise linear models were investigated for the continuous thickness of layer 2. Unfortunately, like their layer 1 counterparts, those models constructed did not show good predictive ability.

Layer 2 thickness was then considered as a two category variable with levels thin/thick, where thin was defined by having a layer 2 thickness < 0.60 metres.

Thickness	NSW	QLD	SA	TAS	VIC	WA	CSIRO
thin ($< 0.60m$)	8506	14891	11944	2416	899	12594	696
thick ($\geq 0.60m$)	6820	28167	5990	1055	891	6155	915
total	15326	43058	17934	3471	1790	18749	1611

Table 60: Layer 2 thin/thick by State.

A *C5.0* classification model was fitted to these data. 34 variables were used; 11 climate, 3 MSS, 16 terrain, lithology, landuse, ASC and the predicted layer 2 thickness as derived from a polyon method (via PPF derived look-up table). The initial model was constructed with a train:test split of 70:30. The performance on the test data yields the following confusion matrix.

		<i>Predicted</i>	
		thin	thick
<i>Observed</i>	thin	9838	4888
	thick	4640	9550

Table 61: Observed versus predicted for two category thickness in layer 2

There are 28916 observations in the test data. The overall classification rate is 0.670. The population specific classification rates are 0.668 for thinner and 0.673 for the thicker layer 2 thickness measurements.

The model was then refitted using the same model options and variables on all layer 2 observations. 49 rules were used in the model.

State	overall		thin		thick	
	N	c/rate	N	c/rate	N	c/rate
NSW	15326	0.60	8506	0.78	6820	0.38
QLD	42405	0.68	14570	0.19	27835	0.93
SA	17295	0.70	11570	0.92	5725	0.25
TAS	3344	0.71	2336	0.95	1008	0.17
VIC	1790	0.58	899	0.88	891	0.28
WA	17892	0.72	12085	0.90	5807	0.34
CSIRO	1573	0.59	681	0.68	892	0.52

Table 62: Classification rate by State for layer 2 thickness (binary thickness model).

Region	overall		thin		thick	
	N	c/rate	N	c/rate	N	c/rate
nterritory	506	0.54	270	0.26	236	0.87
carpentaria	205	0.47	111	0.43	94	0.51
qldnorth	563	0.56	274	0.28	289	0.82
qldcentral	10699	0.72	3147	0.22	7552	0.93
qldsouth	10969	0.62	4408	0.14	6561	0.94
moreton	5638	0.63	2115	0.12	3523	0.94
burdekin	7667	0.69	2621	0.32	5046	0.88
fitzroy	6041	0.73	1820	0.17	4221	0.97
mdb	13656	0.65	7429	0.85	6227	0.41
nswnorth	2247	0.58	1048	0.27	1199	0.86
nswsouth	5838	0.60	3393	0.89	2445	0.20
viceast	842	0.61	516	0.94	326	0.09
vicwest	2155	0.72	1567	0.99	588	0.02
tasmania	3343	0.71	2335	0.95	1008	0.17
sagulf	10915	0.69	7120	0.91	3795	0.29
wplateau	593	0.71	412	0.93	180	0.22
wasouth	16293	0.72	11285	0.92	5008	0.29
indian	1456	0.62	776	0.66	680	0.58

Table 63: Classification rate by region for layer 2 thickness (binary thickness model).

Tables 62 and 63 give the classification rates, both overall and for classes thin and thick for all States and regions.

The classification rates for thick and thin are at times quite unbalanced. For example in Tasmania 95% of the 2335 thin soils are classified correctly but only 17% of the 1008 thicker soils.

Tables 62 and 63 suggest that in Queensland and northern New South Wales (not MDB) the model is notably stronger at identifying thicker layer 2 soils, while elsewhere the thinner soils are better predicted. Thicker layer 2 soils are particularly poorly predicted in Victoria and to a lesser extent Tasmania, while thinner layer 2 soils are not well predicted in southern Queensland.

The 49 rules from the final *C5.0* model were applied to ASRIS extent to generate a map of layer 2 thickness class predictions. These predictions can be viewed at www.nlwra.gov.au/data.

An associated model certainty surface was created according to the method described in Section 3.8. The environmental representativeness surface incorporated 8 predictor variables, namely: relief, relative elevation, ridge distance, MSS band 4, highest period moisture index, annual mean radiation, annual mean precipitation and the maximum temperature of the warmest period. This certainty surface is available at www.nlwra.gov.au/data.

11 APPENDIX: pH IN WATER TO pH IN CaCl_2 : A CALIBRATION EQUATION

The two most common measurements of soil pH made in Australian laboratories are in water (pH_w) and in CaCl_2 (pH_c), both typically at a soil to solution ratio of 1:5. Various methods for calibrating pH_c and pH_w have been considered in the literature as there are often times when one needs to change from one assessment to the other. Slattery et al. (1999) summarize the existing calibration equations. All the reported equations are linear, with the largest sample size used being 7844 from Ahern et al. (1995). Slattery et al. (1999) do however remark that while the calibration functions are linear, there are a number of authors who have shown that this relationship is not linear over the entire range, citing the articles of Aitken & Moody (1991), Little (1992) and Ahern et al. (1995).

The purpose of this appendix is to obtain a more definitive assessment of the relationship between pH_c and pH_w through investigation of a much larger data set than previously considered.

The data used to derive a new pH water to pH CaCl_2 calibration equation comes from two sources. The first is data from NLWRA Theme 5, Project 4D on “Nutrient balance in regional farming systems and soil nutrient status”, where there are 58548 estimates of pH in both CaCl_2 and water from South Australia available. These measurements are surface measurements, the vast majority 0-10 centimetres.

The second source is the ASRIS database where there are some 11917 dual assessments available from throughout the country and across all horizon layers. 3083 of these dual assessments are from the first layer.

These data sets can be combined to give a total of 70465 observations with both pH water and pH CaCl_2 assessments. No account was made for depth as the intention was to derive a single calibration equation broadly applicable to all depths.

Polynomial calibration models were examined in the first instance for their ease of application. While they performed well within the central section of the data, their tail behaviour was considered unreasonable.

An additive model (Hastie & Tibshirani, 1990) was then trialed as an alternative. This is a generalization of the familiar linear regression model where by smooth functions of the individual predictors are allowed. This is typically written as

$$Y = \alpha + \sum_{i=1}^p f_j(X_j) + \epsilon$$

where f_j denotes a some smooth function of predictor X_j and the errors ϵ are assumed to be independent with mean zero and constant variance σ^2 .

In our context, pH_c is the response and pH_w is the obvious predictor variable given our desire to update pH_w to pH_c . An additive model was fitted with the smooth function of pH water created by a smoothing spline with 6 degrees of freedom. The choice of the degrees of freedom was made after examining the generalized cross validation score as a function of the degrees of freedom. Beyond 6 degrees of freedom the increasing complexity did not appear to offer any gains to the modelling.

The fitted additive model is superimposed on a random sample of 5000 points from the combined data set in Figure 69. The additive model appears to be stable outside the range of the data and performs well in summarizing the non linear relationship between pH_w

and pH_c ($R^2 = 96.2$, $\text{MSE} = 0.058$). The additive model reduces the deviance from the linear model by 869.5 on the 5 additional degrees of freedom.

4A1 vs 4B1

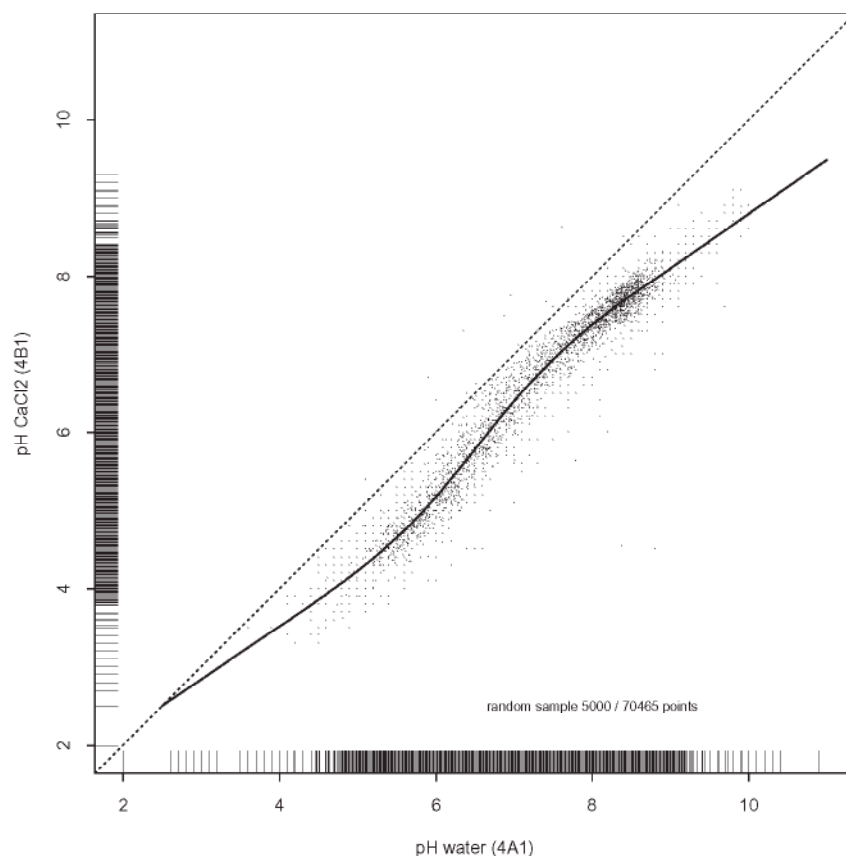


Figure 69: Additive model: pH in CaCl_2 versus pH in water.

The residuals from the model show no obvious unexplained structure. Several large outliers can be identified in the random sample of the 5000 points in Figure 69 as they have markedly lower pH_c for their associated pH_w values.

The predictions equations for additive models can not be written down in an analogous way to polynomial models. The predictions are however summarised in a look up table, Table 64.

Observed pH water	Predicted pH CaCl ₂	Observed pH water	Predicted pH CaCl ₂	Observed pH water	Predicted pH CaCl ₂
2.4	2.4	5.0	4.2	7.6	7.0
2.5	2.5	5.1	4.3	7.7	7.1
2.6	2.6	5.2	4.4	7.8	7.2
2.7	2.6	5.3	4.5	7.9	7.3
2.8	2.7	5.4	4.6	8.0	7.4
2.9	2.8	5.5	4.7	8.1	7.5
3.0	2.8	5.6	4.8	8.2	7.5
3.1	2.9	5.7	4.9	8.3	7.6
3.2	3.0	5.8	5.0	8.4	7.7
3.3	3.0	5.9	5.1	8.5	7.8
3.4	3.1	6.0	5.2	8.6	7.8
3.5	3.2	6.1	5.3	8.7	7.9
3.6	3.2	6.2	5.4	8.8	8.0
3.7	3.3	6.3	5.5	8.9	8.0
3.8	3.4	6.4	5.7	9.0	8.1
3.9	3.5	6.5	5.8	9.1	8.2
4.0	3.5	6.6	5.9	9.2	8.2
4.1	3.6	6.7	6.0	9.3	8.3
4.2	3.7	6.8	6.2	9.4	8.4
4.3	3.7	6.9	6.3	9.5	8.4
4.4	3.8	7.0	6.4	9.6	8.5
4.5	3.9	7.1	6.5	9.7	8.6
4.6	3.9	7.2	6.6	9.8	8.7
4.7	4.0	7.3	6.7	9.9	8.7
4.8	4.1	7.4	6.8	10.0	8.8
4.9	4.2	7.5	6.9	10.1	8.9

Table 64: pH water to pH CaCl₂ look up table.

Other potential predictors were considered. Most notably, organic carbon and total nitrogen were considered as additional predictors, both in continuous and categorical form. They were however found to have negligible explanatory power.

REFERENCES

- AHERN, C. R., BAKER, D. E., & AITKEN, R. L. (1995). Models for relating pH measurements in water and calcium chloride for a wide range of pH, soil types and depths. *Plant and Soil* **171**, 47–52.
- AITKEN, R. L. & MOODY, P. W. (1991). Interrelations between soil pH measurements in various electrolytes and soil solution pH in acidic soils. *Australian Journal of Soil Research* **29**, 483–491.
- BECKETT, P. H. T. & WEBSTER, R. (1971). Soil variability: a review. *Soils and Fertilizers* **34**, 1–15.
- BREIMAN, L., FRIEDMAN, J. H., OLHSEN, R. A., & STONE, C. J. (1984). *Classification and Regression Trees*. Wadsworth, Monterey, California.
- BUREAU OF RURAL SCIENCES (2000). Digital Atlas of Australian Soils. www.affa.gov.au/docs/rural_science/datasets/atlas/index.html.
- CARLILE, P., BUI, E., MORAN, C., SIMON, D., & HENDERSON, B. (2001). Method used to generate soil attribute surfaces for ASRIS using soil maps and look up tables. Technical Report 24/01, CSIRO Land and Water.
- COOK, S. E., CORNER, R. J., GREALISH, G., GESSLER, P. E., & CHARTRES, C. J. (1996). A rule-based system to map soil properties. *Soil Science Society of America Journal* **60**, 1893–1900.
- DE GRUITJER, J. J. (2000). Sampling for spatial inventory and monitoring of natural resource. Technical report, Alterra-rapport 070, Wageningen, Alterra, Green World Research.
- FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines. *The Annals of Statistics* **19**, 1–141.
- GESSLER, P. E. (1996). *Statistical soil-landscape modelling for environmental management*. PhD thesis, Australian National University, Canberra.
- GESSLER, P. E., MOORE, I. D., MCKENZIE, N. J., & RYAN, P. J. (1995). Soil-landscape modelling and spatial prediction of soil attributes. *International Journal of Geographical Information Systems* **4**, 421–432.
- HASTIE, T. J. & TIBSHIRANI, R. J. (1990). *Generalized additive models*. Chapman and Hall, London.
- HEANES, D. L. (1984). Determination of total organic-C in soils by an improved chromic digestion and spectrophotometric procedure. *Communications in Soil Science and Plant Analysis* pages 1191–1213.
- HUDSON, B. D. (1992). The soil survey as a paradigm-based science. *Soil Science Society of America Journal* **56**, 836–841.
- JENNY, H. (1941). *Factors of soil formation*. McGraw-Hill, New York.
- JENNY, H. & LEONARD, C. D. (1934). Functional relationships between soil properties and rainfall. *Soil Science* **38**, 363–381.
- KIIVERI, H. & CACCETTA, P. (1998). Image fusion with conditional probability networks for monitoring the salinization of farmland. *Digital signal processing* **8**, 225–230.
- KREZNOR, W. R., OLSON, K. R., BANWART, W. L., & JOHNSON, D. L. (1989). Soil, landscape, and erosion relationships in a Northwest Illinois watershed. *Soil Science Society of America Journal* **53**, 1763–1771.
- LIM, T. S., LOH, W. Y., & SHIH, Y. S. (2000). A comparison of prediction accuracy, complexity and training time of thirty three old and new classification algorithms. *Machine Learning* **40**, 203–228.
- LITTLE, I. P. (1992). The relationship between soil pH measurements in calcium chloride

- and water suspensions. *Australian Journal of Soil Research* **30**, 587–592.
- MCKENZIE, N. J., JACQUIER, D., ASHTON, L. J., & CRESSWELL, H. P. (2000). Estimating soil properties using the Atlas of Australian Soils. Technical Report 11, CSIRO Land and Water, Canberra.
- MCKENZIE, N. J. & RYAN, P. J. (1999). Spatial prediction of soil properties using environmental correlation. *Geoderma* **89**, 67–94.
- MOORE, I. D., GESSLER, P. E., NIELSEN, G. A., & PETERSON, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal* **57**, 443–452.
- NORTHCOTE, K. H., BECKMANN, G. G., BETTENAY, E., CHURCHWARD, H. M., DIJK, D. C. V., DIMMOCK, G. M., HUBBLE, G. D., ISBELL, R. F., MCARTHUR, W. M., MURTHA, G. G., NICOLLS, K. D., PATON, R., THOMPSON, C. H., WEBB, A. A., & WRIGHT, M. J. (1960-1968). *Atlas of Australian Soils. Sheets 1-10, with explanatory booklets*. CSIRO and Melbourne University Press, Melbourne.
- ODEH, I. O., MCBRATNEY, A. B., & CHITTLEBOROUGH, D. J. (1994). Spatial prediction from land form attributes derived from a digital elevation model. *Geoderma* **63**, 197–214.
- QUINLAN, J. R. (1992). Learning with continuous classes. In *Proceedings of the 5th Australian Joint conference on artificial intelligence*, pages 343–348. World Scientific.
- QUINLAN, J. R. (1993a). *C4.5: Programs for machine learning*. Morgan-Kaufman, San Mateo, California.
- QUINLAN, J. R. (1993b). Combining instance-based and model-based learning. In *Proceedings of the 10th International conference on machine learning*, pages 236–243, San Francisco.
- RAYMENT, G. E. & HIGGINSON, F. R. (1992). *Australian laboratory handbook of soil and water chemical methods*. Inkata Press, Melbourne.
- SKJEMSTAD, J. O., SPOUNCER, L. R., & BEECH, A. (2000). Carbon conversion factors for historical soil carbon data. Technical Report 15, National Carbon Accounting System, Australian Greenhouse Office.
- SLATTERY, W. J., CONYERS, M. K., & AITKEN, R. L. (1999). *Soil Analysis: an interpretation manual*, chapter Soil pH, aluminium, manganese and lime requirement. CSIRO Publishing, Melbourne.
- SPEIGHT, J. G. (1974). A parametric approach to landform regions. *Special publication Institute of British Geographers* **7**, 213–230.
- THERNEAU, T. M. & ATKINSON, E. J. (1997). An introduction to recursive partitioning using the RPART routine. Technical Report 61, Mayo Clinic, Section of Statistics.
- UYSAL, I. & GÜVENİR, H. A. (1999). An overview of regression techniques for knowledge discovery. *Knowledge Engineering Review* **14**, 319–340.
- WALKER, P. H., HALL, G. F., & PROTZ, R. (1968). Relationship between landform parameters and soil properties. *Soil Science Society of America Proceedings* **32**, 101–104.
- WALKLEY, A. (1947). A critical examination of a rapid method for determining organic carbon in soils - effect of variations in digestion conditions and of inorganic constituents. *Soil Science* **63**, 251–264.