7. Benchmark Examples

In this section we present tree-based profiling and clustering methods on three benchmark datasets. Each of these datasets has been selected to highlight features of tree-based methods and to compare their performances. The datasets selected are all freely available benchmark datasets.

The first dataset is the Thyroid dataset, which is a clustering problem involving only quantitative variables. Here the improvement in clustering performance gained through using the auto-association proximity matrices is shown. MCTs are compared with auto-associative random forests and treeboost, AA-MRTs, PAM and K-means.

The second dataset is the Wisconsin breast cancer dataset. This analysis is used to compare the performance of tree methods in a categorical domain with a known clear grouping structure. For this analysis MCT approaches are compared to binary substitution and Gower distance methods.

The third dataset is the horse colic dataset. This analysis is focused on the performance on MCTs in a mixed domain profiling problem. Here the limits of the Gower distance and binary substitution methods are shown and the power of the proximity matrices is highlighted. This study also explores the features of MCTs that assist in further understanding and simplifying the problem. In particular the ability of PLAID consensus generation to find subgroups within variables of the profiling set is highlighted.

7.1 Clustering Quantitative Variables: Thyroid dataset

In this analysis a comparison between tree-based methods for clustering and existing methods is performed using the thyroid dataset. The thyroid dataset (Coomans, Broeckaert, Jonckheer and Massart 1983) consists of 215 observations on 5 variables that describe the action of the thyroid gland. There are three known groups in the data corresponding to hypothyroid (1), hyperthyroid (2) and normal (3) patients. The other variables are hormone levels measured in the blood. These are:

1. TSH
2. DTSH
3. RT3U
4. T4
5. T3

The goal of the analysis is to cluster the data and compare the clustering performance with the known groups.

7.1.1 AA-MRT

The RE graph for AA-MRT is used to determine the size of the tree (Figure 50), and it can be seen that the performance plateaus at 5 terminal nodes, and the corresponding tree is displayed in Figure 51. Using the terminal node locations as group classifications, AA-MRTs misclassify 35 observations (Table 18) when compared with the known groups. The AA-MRT of the raw data outperforms PAM, which misclassified 49 observations, but not K-means, which misclassified 30 observations.

Figure 50: Thyroid analysis AA-MRT RE graph.

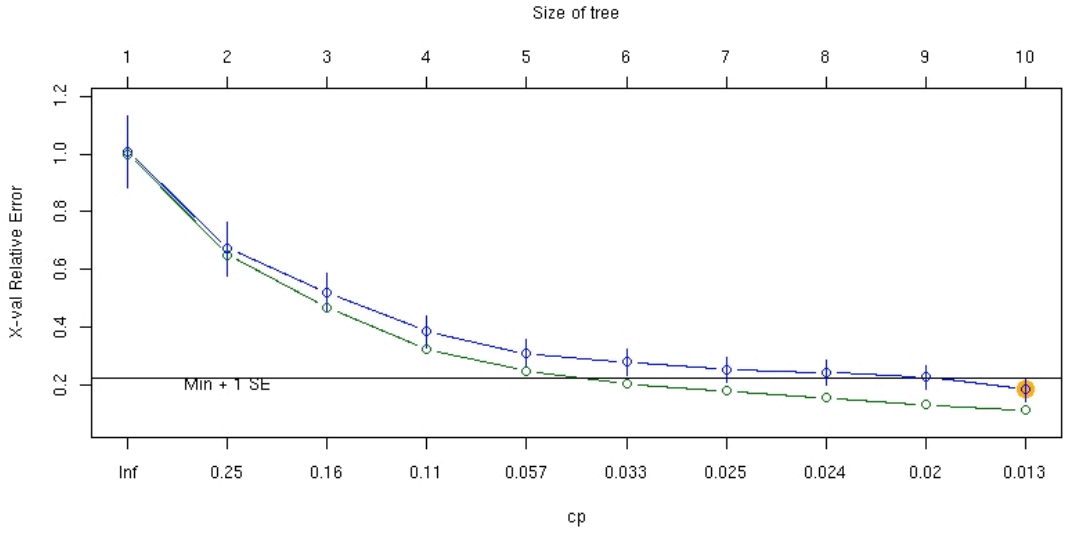


Figure 51: Thyroid analysis AA-MRT.

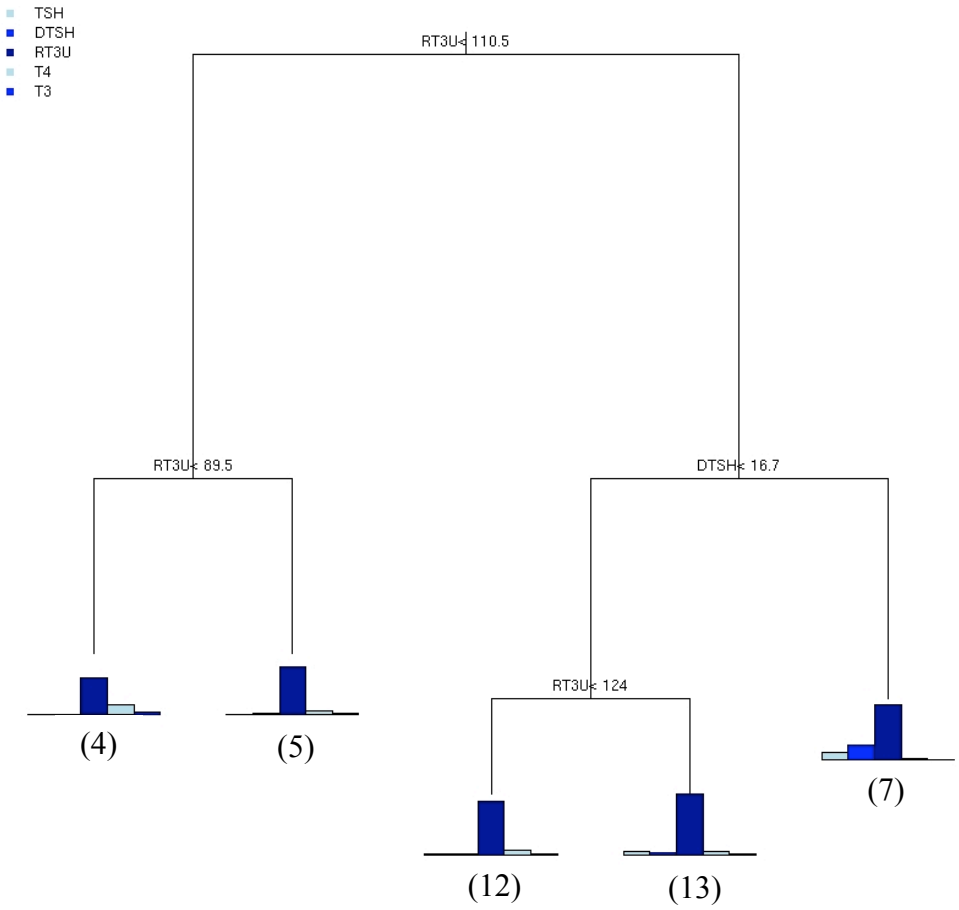


Table 18: Thyroid analysis AA-MRT misclassification table

	Terminal Node				
	4	**5**	**7**	**12**	**13**
Hypothyroid	0	4	4	10	12
Hyperthyroid	16	14	2	3	0
Normal	0	78	64	8	0

7.1.2 AA-RF

AA-RF performs quite well on the dataset, converging to a stable predictive performance after 100 trees are added to the model (Figure 52). This precision is mirrored within the proximity matrix and MDS images with three groups obvious (Figure 53). Clustering this matrix with an MCT using SSR splitting (Figure 54),

gives 22 misclassifications (Table 19) and 12 misclassifications are recorded for PAM and 11 for K-means. It is clear that the clustering techniques all do better on the proximities than on the raw data, whereas the fact that K-Means and PAM do better than MCTs is a reflection on the overlapping nature of the groups. If the groups are strongly overlapping it is likely that a partition on a single variable will be sufficient. Both K-means and PAM have the luxury of not requiring a clear single variable separation between the groups and therefore do better.

Figure 52: Thyroid analysis AA-RF error convergence plot.

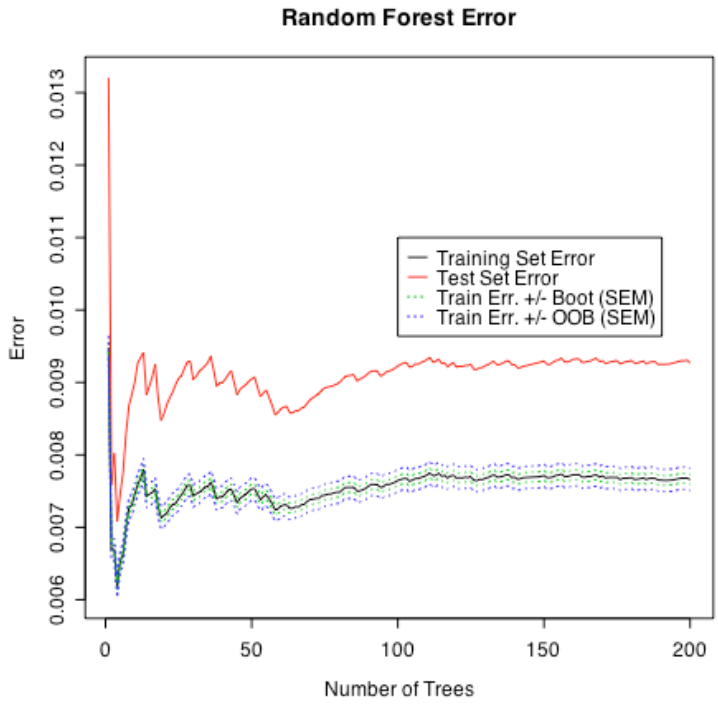


Figure 53: Thyroid analysis AA-RF proximity images.

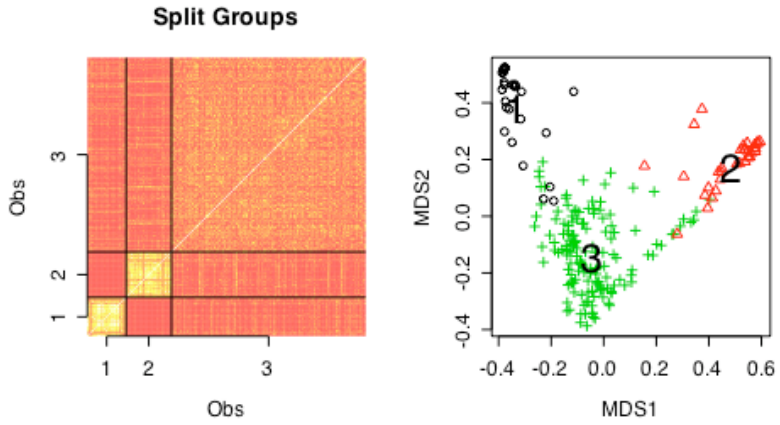


Figure 54: Thyroid analysis SSR partition on the AA-RF proximity matrix.

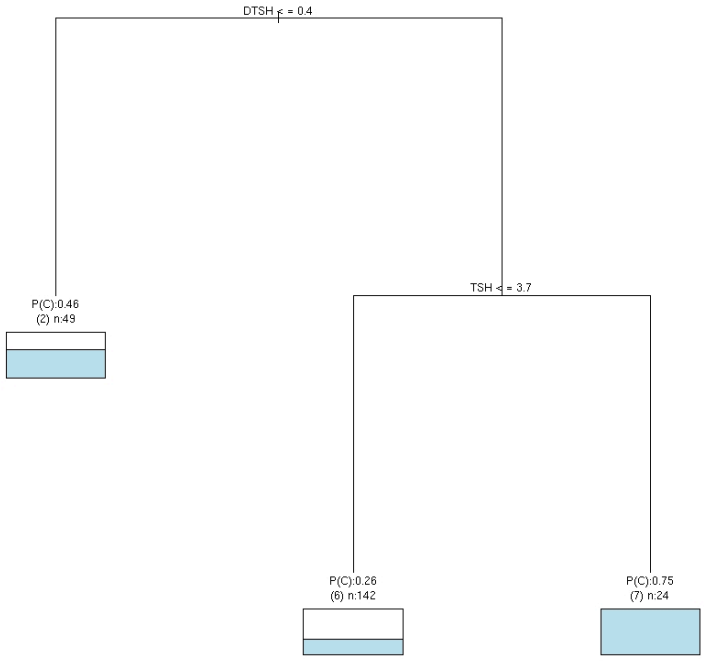


Table 19: Thyroid analysis AA-RF misclassification table

	Terminal Node		
	2	**6**	**7**
Hypothyroid	0	34	15
Hyperthyroid	6	1	135
Normal	24	0	0

7.1.3 AA-Treeboost

The AA-Treeboost model proved to be more complex than the random forest models with the error converging (Figure 55) after 300 trees were added to the model. Despite the number of trees added the proximity images do not obviously show three known groups (Figure 56). This non-obvious structure affects the performances of the base clustering algorithms with K-Means and PAM misclassifying 38 observations. However MCTs with MR splitting (Figure 57) on the treeboost proximity matrix misclassified only 15 observations (Table 20). The improvement gained by MCTs is most likely a direct result of trees being used to construct the proximity matrix, thus allowing MCTs to identify the structure not easily found by other methods.

Figure 55: Thyroid analysis AA-Treeboost error convergence plot.

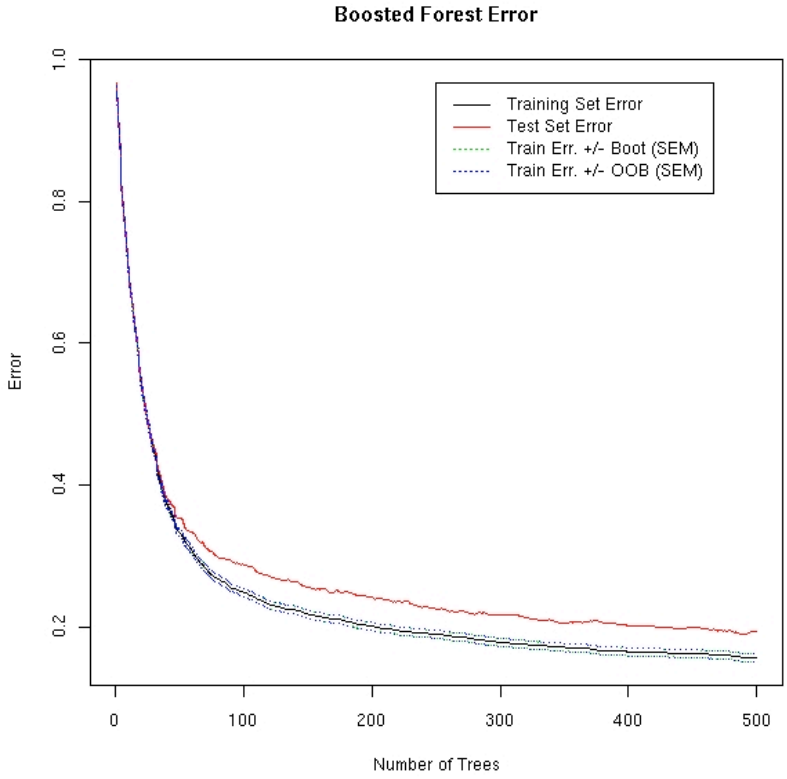


Figure 56: Thyroid analysis AA-Treeboost proximity images.

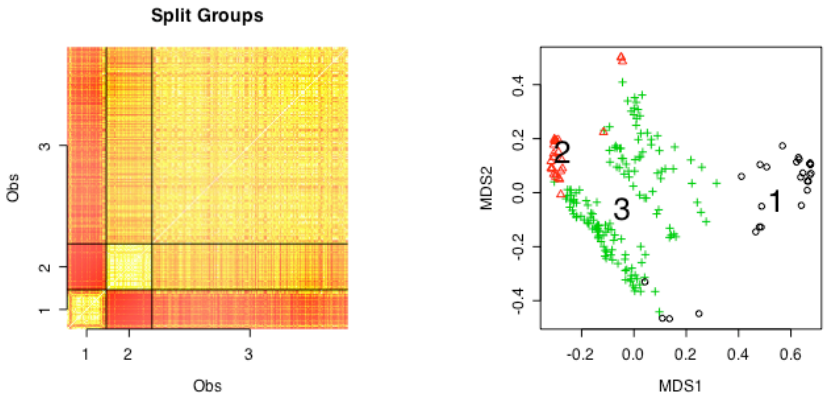


Figure 57: Thyroid analysis MR partition on the AA-Treeboost proximity matrix.

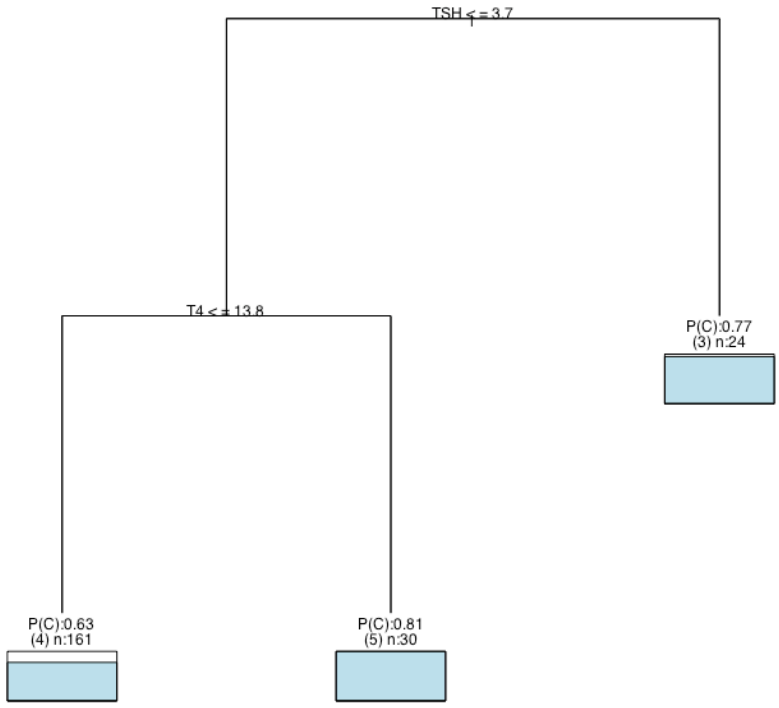


Table 20: Thyroid analysis AA-Treeboost misclassification table.

	Terminal Node		
	3	**4**	**5**
Hypothyroid	24	6	0
Hyperthyroid	0	7	28
Normal	0	148	2

7.1.4 Global MCT

The cross-validation for Global MCTs using SSR splitting identifies 2 splits or three groups (Figure 58). The proximity matrix images (Figure 59) are comparable to those found by AA-RF. The corresponding MCT (Figure 60) misclassifies 19 observations (Table 21). However by observation of each terminal node's probability of expression, "P(C)" it can be seen that node 3, which corresponds to the normal group is under-expressed showing a probability of 0.32. This implies that this group is difficult for trees to correctly classify. A finding that is mirrored by its broad dispersion over the MDS plot (Figure 59). When compared with K-means and PAM, MCTs are found to under-perform as they both only misclassify 9 observations.

Figure 58: Thyroid analysis global MCT 10-Fold cross-validated RE curves.

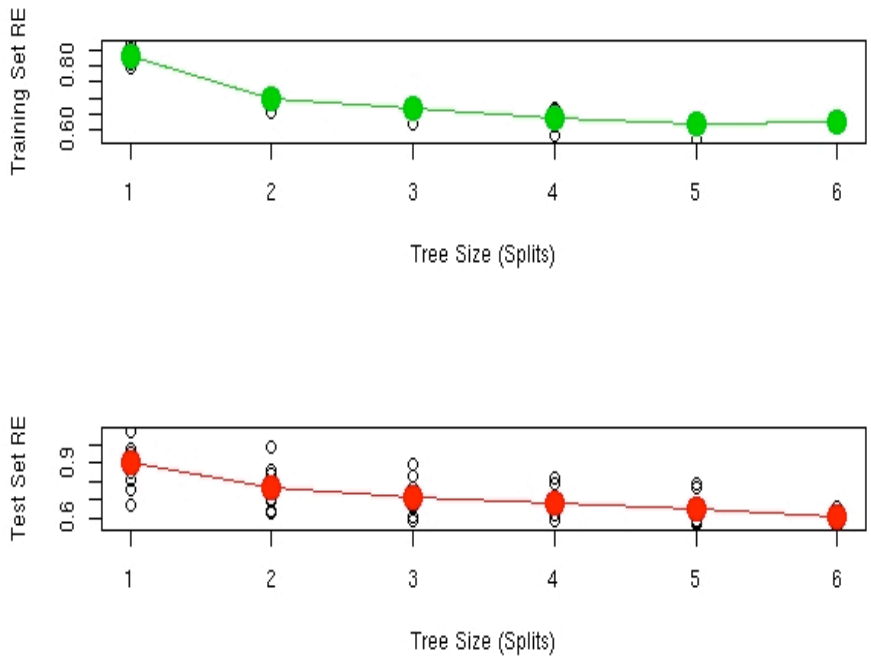


Figure 59: Thyroid analysis global MCT proximity images.

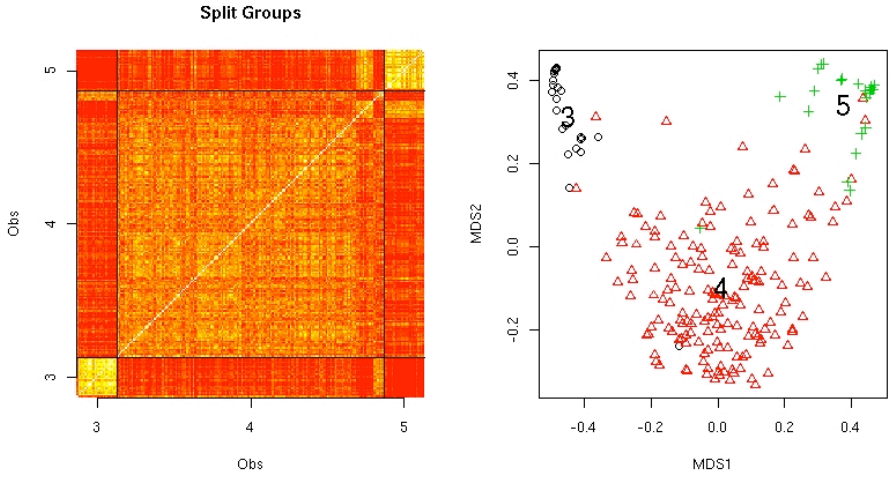


Figure 60: Thyroid analysis global MCT, constructed with MR splitting on the GPA consensus.

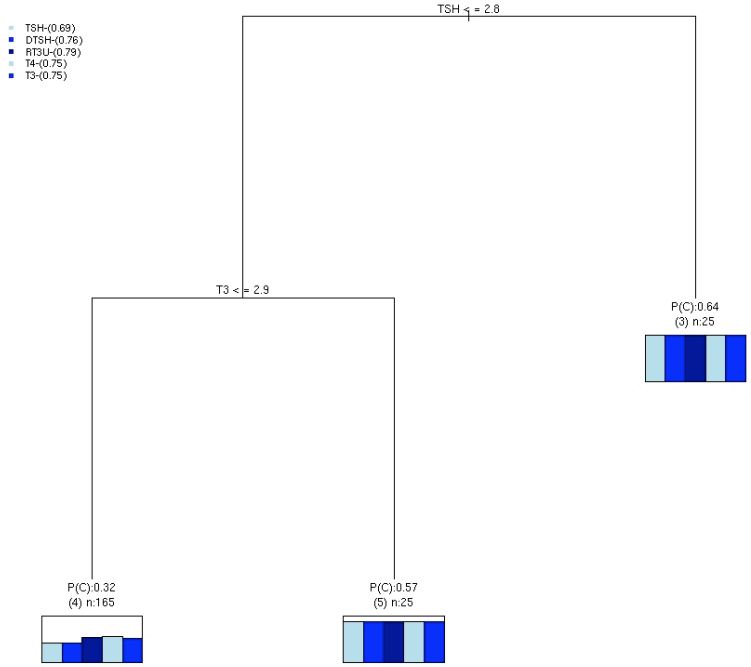


Table 21: Thyroid analysis global MCT misclassification table.

	Terminal Node		
	3	**4**	**5**
Hypothyroid	24	6	0
Hyperthyroid	0	11	24
Normal	1	148	1

7.1.5 Local MCT

Local MCTs find a more complicated tree than global MCTs, identifying 3 splits or 4 groups (Figure 61). These groups are clearly observed within the ACM images and MDS plots, (Figure 62) and this improved resolution is also obvious in the terminal probabilities of the corresponding tree (Figure 63), which are significantly greater than for global MCTs. The local MCT equalled the performance of boosting MCTs, misclassifying 15 observations (Table 22), however the proximity matrices have a more defined structure. However using the ACM, K-means (finding 4 groups) and PAM (finding 3 groups) only misclassified 12 and 10 observations respectively.

Figure 61: Thyroid analysis local MCT RE and AIC plots.

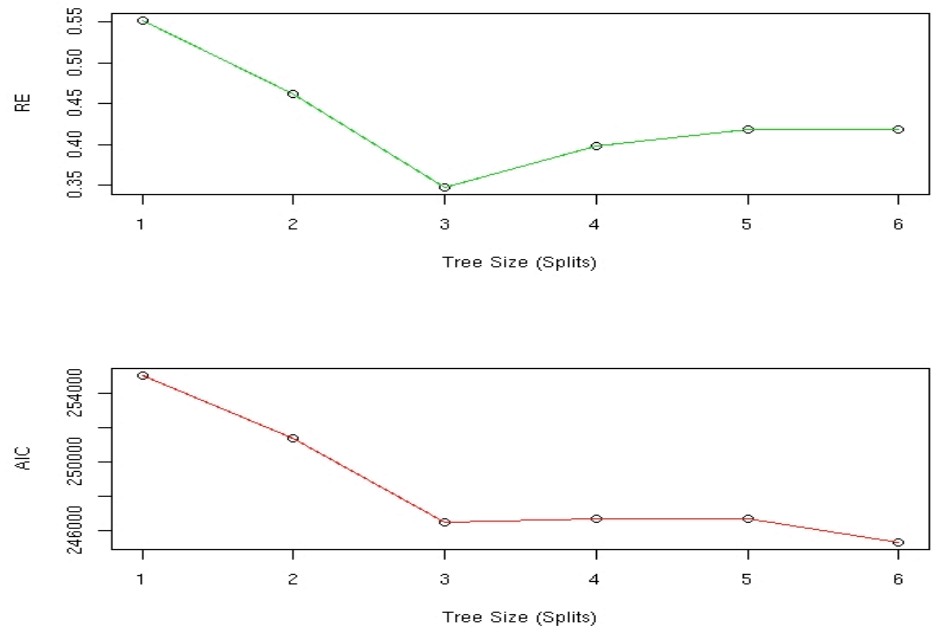


Figure 62: Thyroid analysis local MCT ACM images and MDS plots.

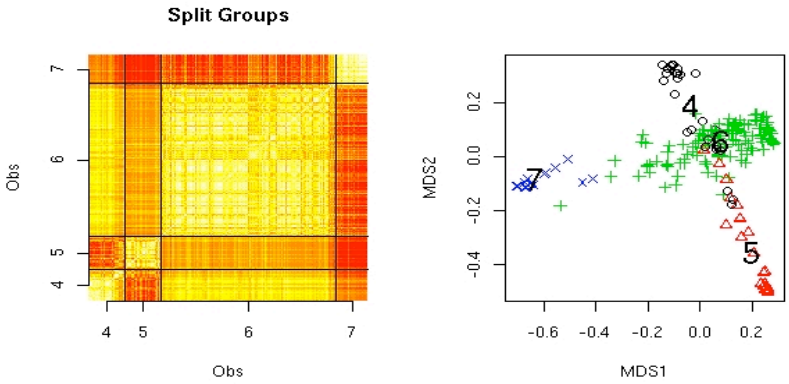


Figure 63: Thyroid analysis local MCT with SSR splitting and GPA consensus combining.

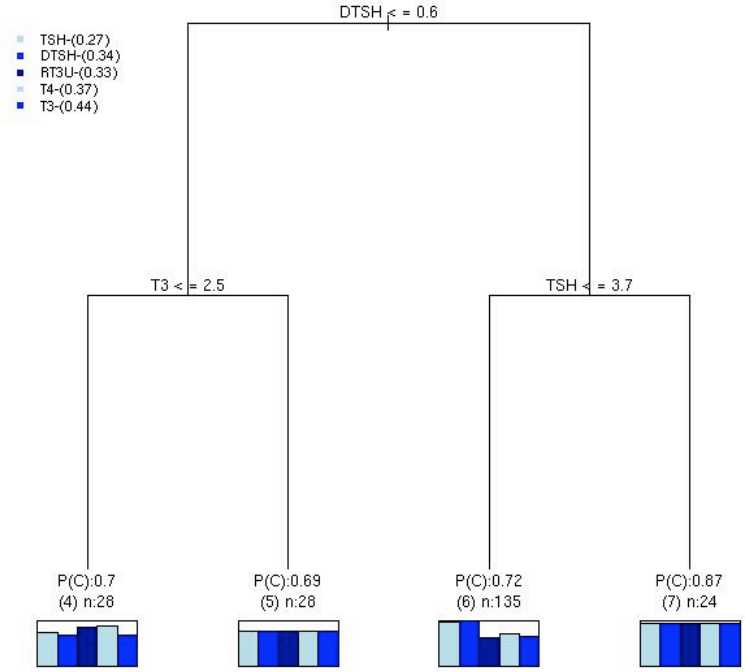


Table 22: Thyroid analysis local MCT misclassification table.

	Terminal Node			
	4	**5**	**6**	**7**
Hypothyroid	0	0	6	24
Hyperthyroid	8	27	0	0
Normal	20	1	129	0

7.1.6 Thyroid summary

A clear result of this example is the marked improvement of general clustering performance that is achieved by using the proximity matrix from either random forests or treeboost. Of the two ensembles the random forest proximity is clearly more stable, and is suited for input into other clustering techniques. The boosted proximity matrix, although producing a more optimal MCT, has a less well defined structure that is not found by other methods.

The fact that K-means and PAM on the proximities do better than trees is primarily due to the fact that trees are constrained by the valid splits available in the variables within the predictor set. When the 15 observations misclassified by MCTs are compared to a classification tree predicting the three groups, which misclassifies 14 observations, it is clear that MCTs are approaching the optimal tree. More so it is clear that the improvements gained by PAM and K-Means are because they are not constrained by the predictor variables.

The differences between the local and global MCTs are expected. Local MCTs have the luxury of removing entire groups, allowing them to focus on groups that may be hard to separate, where as global MCTs are always observing the entire dataset. In this example, the local MCT was more complicated, however more accurate. This accuracy is found not only in the misclassification performances but also in the probability of expression for each terminal node of the tree. Global MCTs found one terminal node that is below random chance expression (3 terminal nodes, random chance expression is P(C)=0.33). By making the tree more complicated the terminal

nodes found by local MCTs were all above random chance expression. As a result the predictive performance of local MCTs is improved.

Overall MCT approaches are shown to improve on AA-MRT, AA-RF and equal the performance of AA-Treeboost. However as they are limited by their tree structure, in this analysis MCTs do not perform as well as PAM or K-means. In fact these methods by searching for groups within the consensus matrix, without knowledge of the known labels, perform better than a classification tree. This highlights the quality of the grouping structure within the consensus and at the same time the limits of a simple tree structured for clustering or classification.

7.2 Clustering Categorical Variables: Breast Cancer Dataset

The breast cancer dataset (Wolberg and Mangasrian 1990) contains 699 observations on 11 variables, one being an index variable, 9 being ordered or nominal, and 1 target class (Table 23). This dataset was sourced from the "*mlbench*" R package (Leisch and Dimitriadou 2005). The aim of this study is to present and compare performances of all tree-based methods for clustering categorical data. For a fair performance comparison the data will be divided in two with 349 training set observations and 350 test set observations.

Firstly the base tree methods are presented. These are Db-MRT on the Gower distance, and MRTs, random forests and treeboost on the binary substituted form of the response. As the response dataset is the binary substituted dataset, these models are not auto-associative. Therefore through this section the random forest and treeboost methods are referred to as binary substituted random forest and binary substituted treeboost. Secondly, the results for local and global MCTs are presented. Finally a summary of the methods and comparison of the results is presented.

Table 23: Breast cancer analysis dataset description.

Variable Name	Description	Type
Id	*Sample code number*	Character
Cl.thickness	*Clump Thickness*	Ordinal {1 to 10}
Cell.size	*Uniformity of Cell Size*	Ordinal {1 to 10}
Cell.shape	*Uniformity of Cell Shape*	Nominal {1 to 10}
Marg.adhesion	*Marginal Adhesion*	Nominal {1 to 10}
Epith.c.size	*Single Epithelial Cell Size*	Ordinal {1 to 10}
Bare.nuclei	*Bare Nuclei*	Ordinal {1 to 10} -16 Missing
Bl.cromatin	*Bland Chromatin*	Nominal {1 to 10}
Normal.nucleoli	*Normal Nucleoli*	Nominal {1 to 10}
Mitoses	*Mitoses*	Nominal {1 to 10}
Class	*Cancer classification*	Nominal {benign, malignant}

7.2.1 Gower dissimilarity Db-MRT

From the RE curve of the Db-MRT (Figure 64) it clear that only two groups have been identified. From the MDS scatter plot (Figure 65) of the distance matrix only two groups found by the tree can be observed. Observation of the misclassification table for the tree in Table 24 show these to groups correspond well with the benign and malignant breast cancers with a misclassification rate of 8 % on the external test set.

Figure 64: Breast cancer analysis Gower distance Db-MRT RE curve.

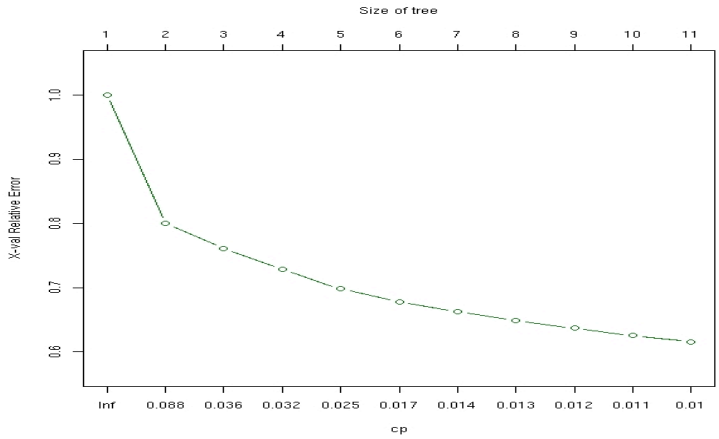
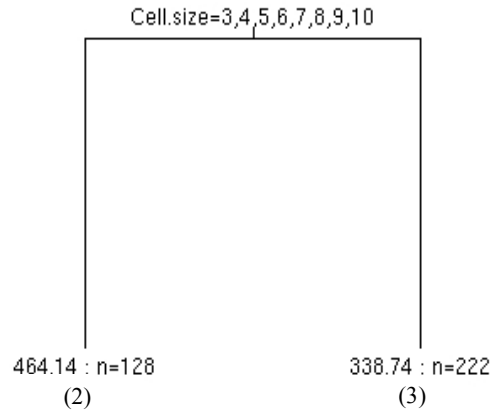
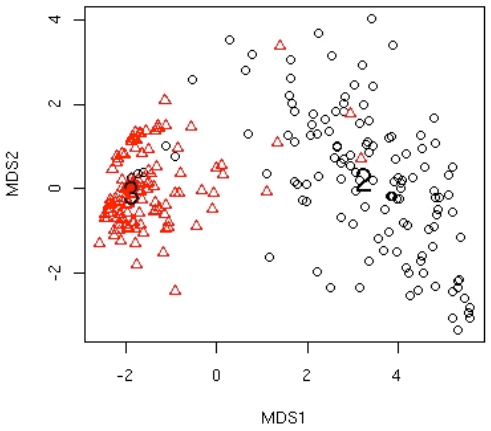


Figure 65: Breast cancer analysis Gower distance Db-MRT.

Gower Distance Db-MRT	MDS terminal node location plot using the Gower Distance Matrix
	

7.2.2 Binary substituted MRT

From the RE curve of the binary substituted MRT (Figure 66) it clear that only two groups have been identified. From the MDS scatter plot (using a Euclidean distance between observations within the response matrix) (Figure 67) only the two groups found by the tree can be observed. Observation of the misclassification table for the tree in Table 24 show these to groups correspond well with the benign and malignant breast cancers with a misclassification rate of 8 % on the external test set.

Figure 66: Breast cancer analysis binary substituted MRT RE curve.

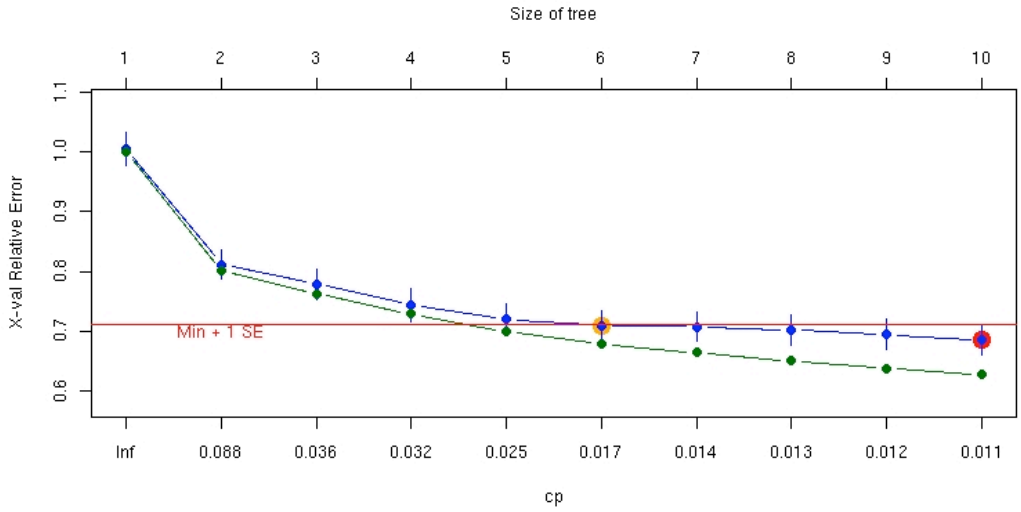
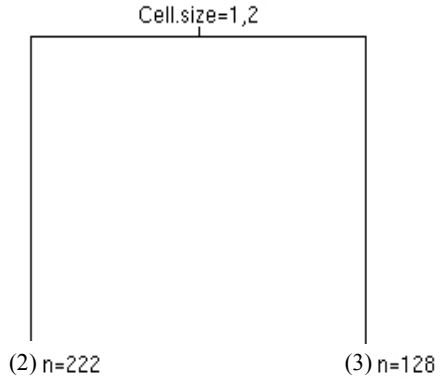
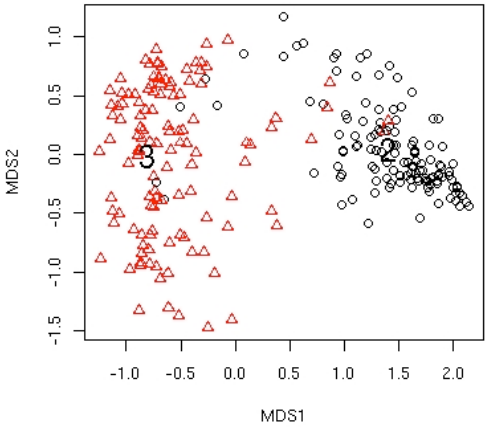


Figure 67: Breast cancer analysis binary substituted MRT and MDS plot.

Binary Substitution MRT	**MDS terminal locations plot using a Euclidean distance over the binary substitution**
	

7.2.3 Binary substituted RF

Binary substituted RF parameters are set to be the following:

(a) A separate random forest test set of 70 training set observations is removed before the analysis to tune the model.

(b) The bootstrapped sample that is used to grow each tree consists of 196 observations and 3 variables.

(c) A maximum tree size of 10 splits within the forest is allowed.

(d) The minimum terminal node size for each tree within the forest is 10 observations.

(e) There are 200 trees within the random forest.

Figure 68: Breast cancer analysis binary substituted AA-RF error convergence plot.

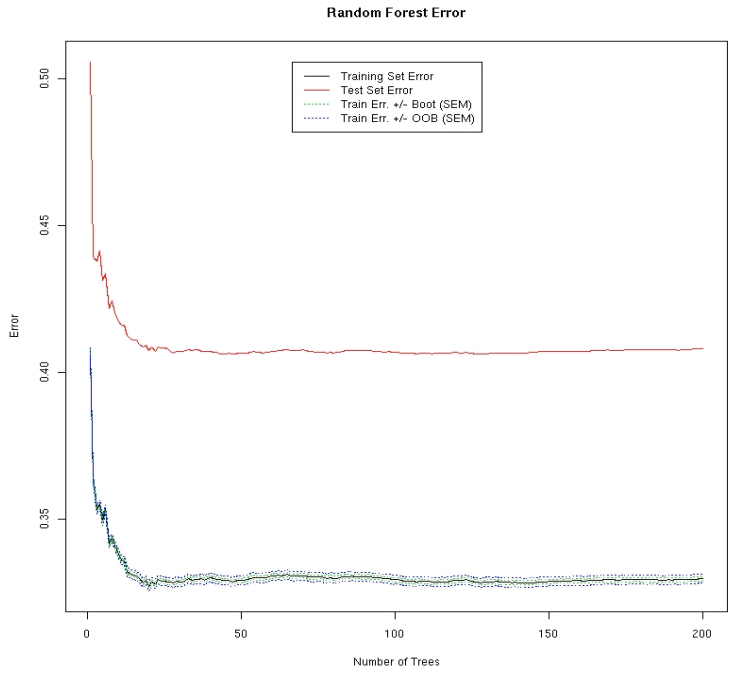


Figure 69: Breast cancer analysis binary substituted RF RE curves.

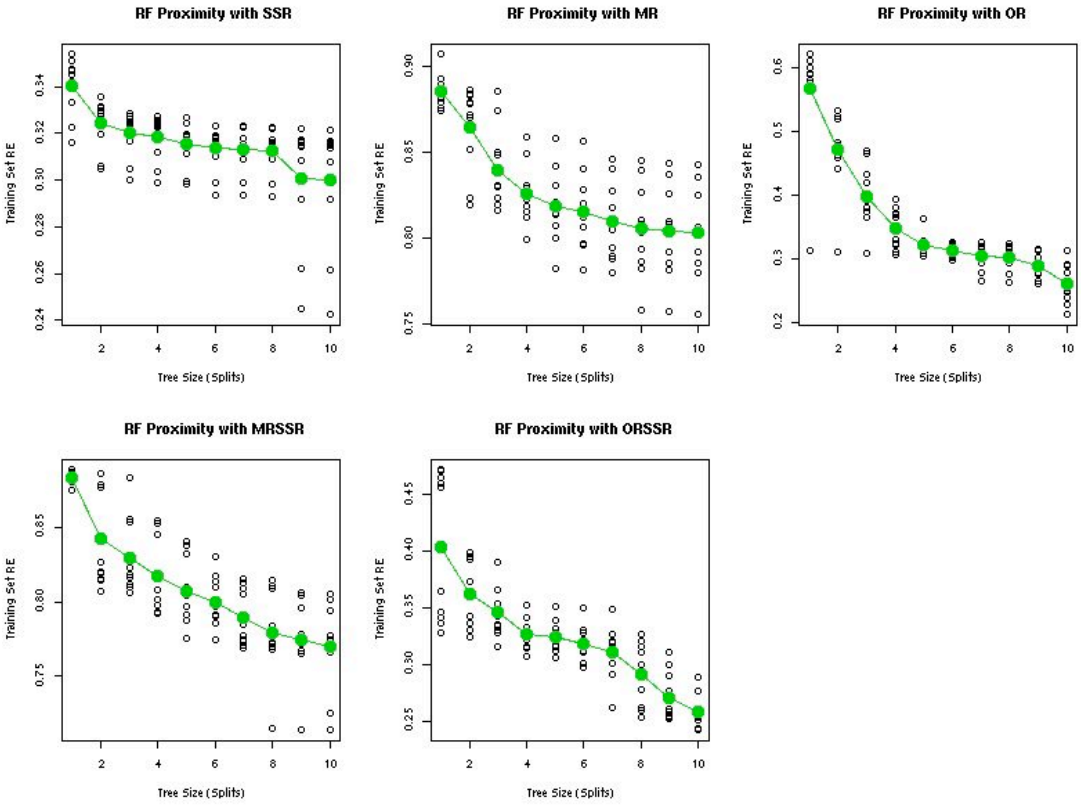


Figure 70: Breast cancer analysis binary substituted random forests MCT built with SSR splitting to 2 splits.

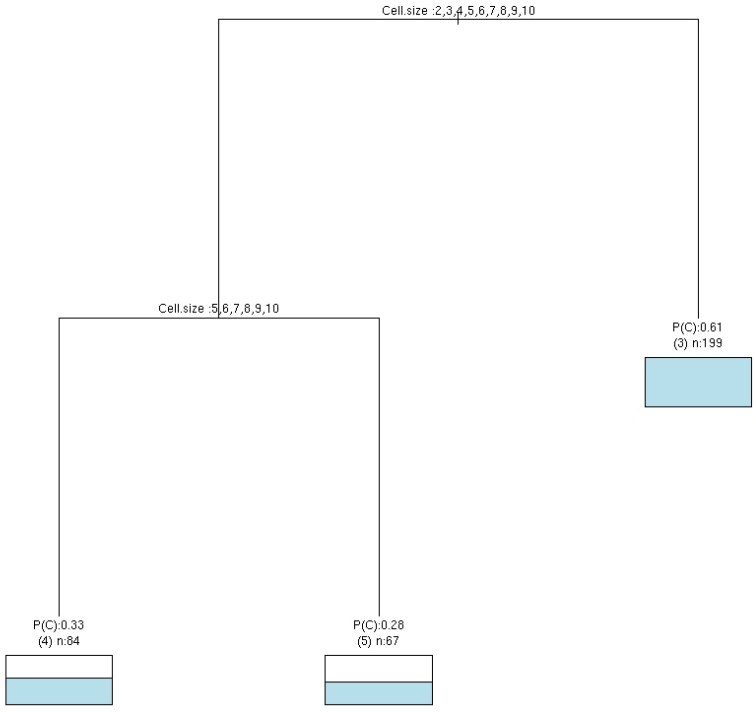
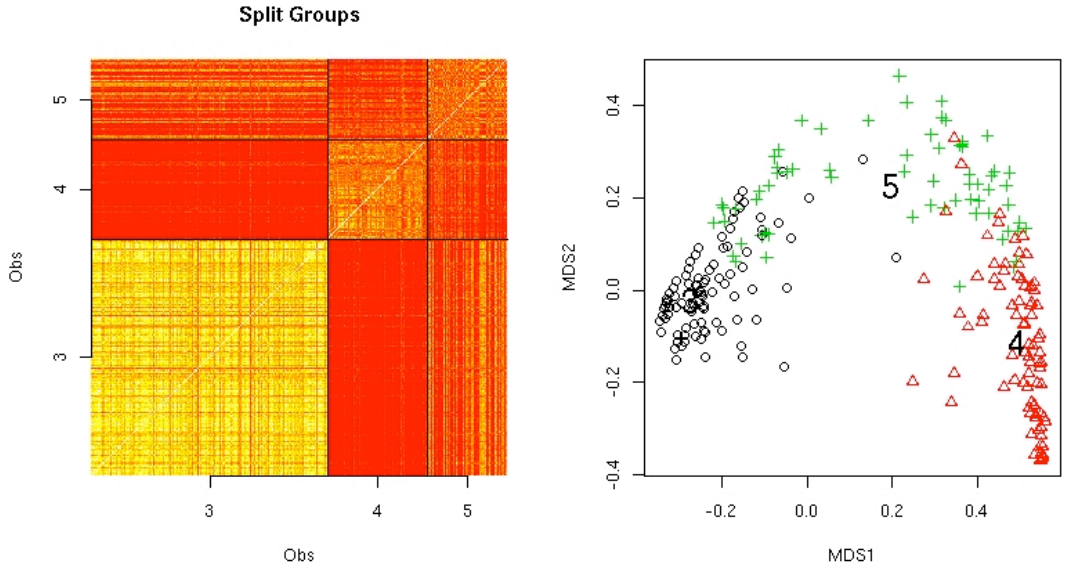


Figure 71: Breast cancer analysis binary substituted RF proximity images.



From the error convergence plot (Figure 68) it is obvious that the random forests error is stable at 200 trees. The RE curves of the MCT splitting criteria however are less clear (Figure 69). Here SSR splitting is selected at two splits, as the RE is stable at approximately 0.32 between 2 and 8 splits. This is not the case with the other splitting criteria. The tree (Figure 70) and the corresponding random forest proximity images (Figure 71) indicate a high certainty in terminal node 3, however markedly less certainty is terminal nodes 4 and 5. This is reflected in the misclassification table (Table 24) with terminal node 3 clearly being the most accurate.

7.2.4 Binary substituted treeboost

The boosted set of trees is grown using the following parameters:

(a) A separate random forest test set of 70 training set observations is removed before the analysis to tune the model.

(b) The bootstrapped sample that is used to grow each tree consists of 196 observations and 3 variables.

(c) A maximum tree size of 2 splits within the boosting is allowed.

(d) The minimum terminal node size for each tree within the boosting is 10 observations.

(e) There are 500 trees within the boosted set.

(f) Shrinkage Parameter set at 0.05.

Figure 72: Breast cancer analysis binary substituted treeboost error convergence plot.

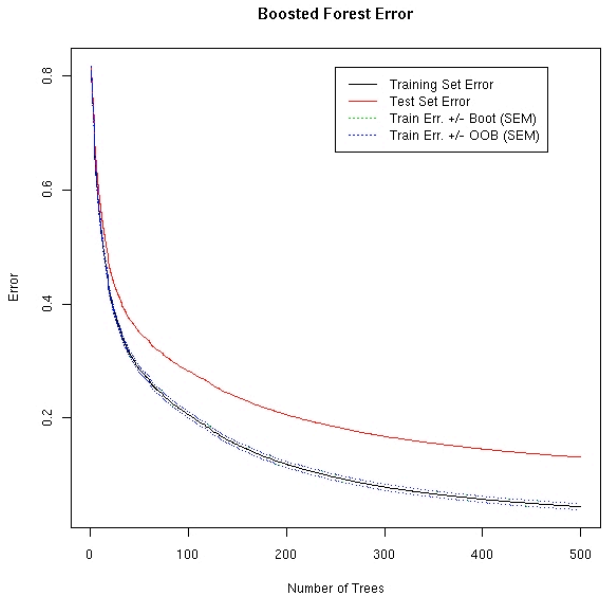


Figure 73: Breast cancer analysis binary substituted treeboost RE curves.

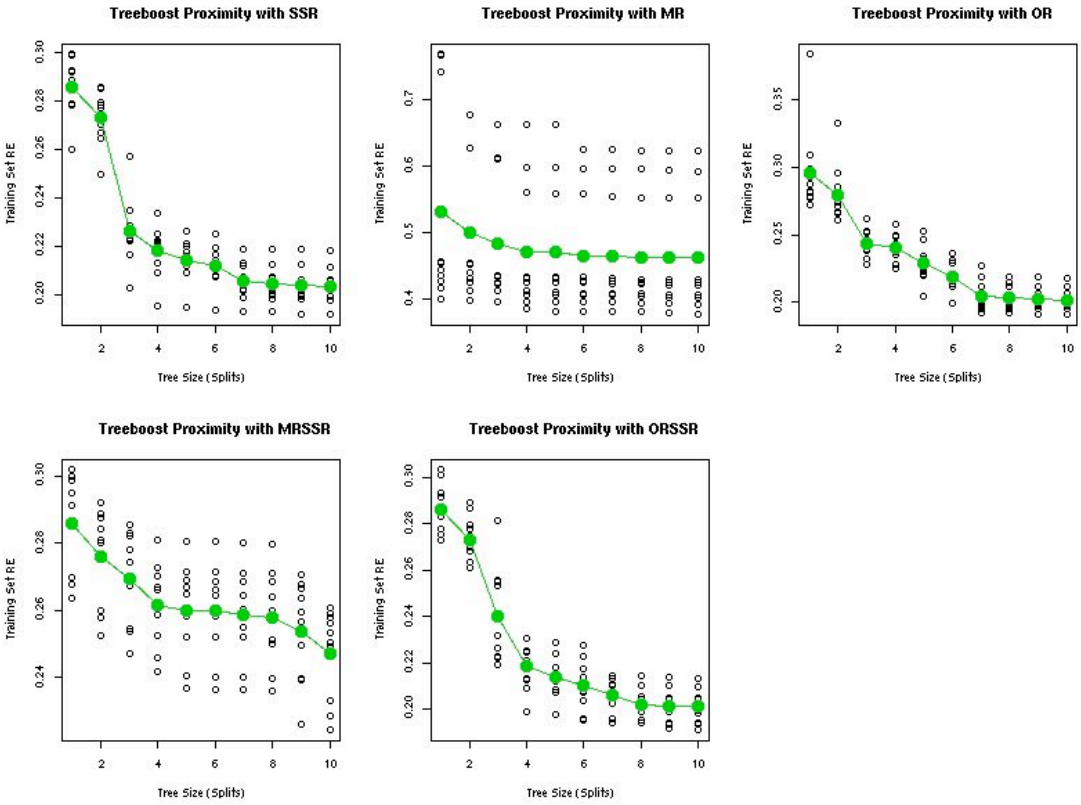


Figure 74: Breast cancer analysis binary substituted treeboost MCT built with SSR splitting to 3 splits.

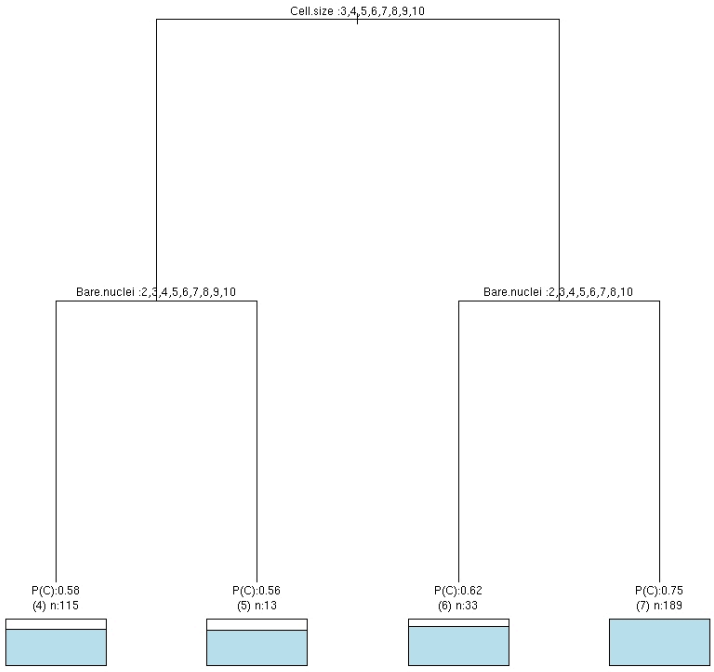
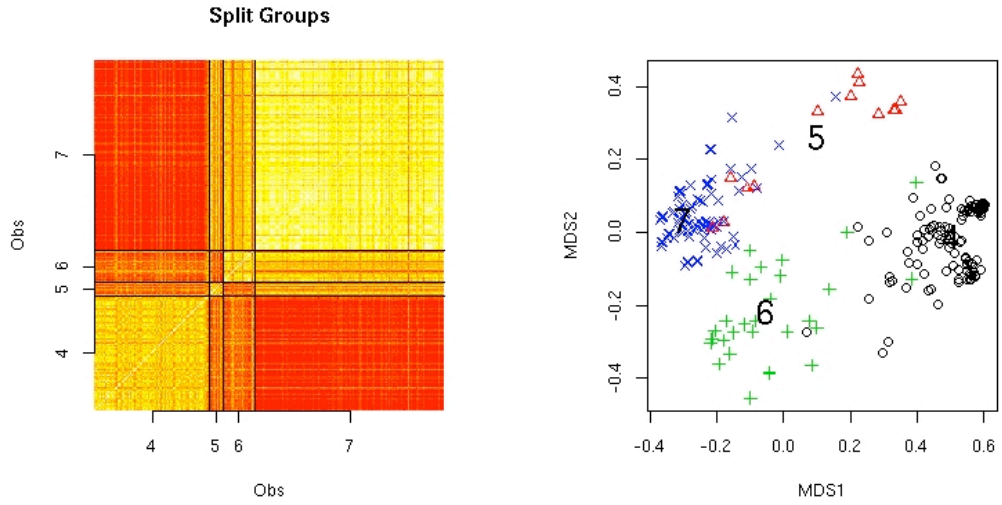


Figure 75: Breast cancer analysis binary substituted treeboost proximity images.



From the error convergence plot (Figure 72) it can be seen that the treeboost model has stabilised after 500 trees. The RE curves (Figure 73) show that the splitting functions SSR, OR and OR-SSR each pick a tree size of 3 splits. Of these SSR is selected, as the cross-validated performances are the most stable at a RE of approximately 0.22. From the tree (Figure 74) and proximity images (Figure 75) a high level of certainty exists throughout each terminal node. This is mirrored in the misclassification table (Table 24), where an error rate of 6.85 % is observed on the test set.

7.2.5 Base method misclassification results

Table 24: Breast cancer analysis misclassification performances of base methods.

Method	Tree Node	Training Set		Test Set	
		Benign	Malignant	Benign	Malignant
Gower DB-MRT (7.5 % Error)	**2**	215	7	202	5
	3	18	110	23	119
	% Error	*7.73 %*	*5.98 %*	*10.36 %*	*3.9 %*
BS-MRT (7.5 % Error)	**2**	7	215	5	202
	3	110	18	119	23
	% Error	*5.98 %*	*7.73 %*	*3.9 %*	*10.36 %*
RF (10.44% Error)	**3**	197	2	183	2
	4	3	81	2	92
	5	33	34	40	30
	% Error	*15.45 %*	*0.89 %*	*0.89 %*	*25.8 %*
Treeboost (6.5 % Error)	**4**	10	105	10	110
	5	8	5	13	9
	6	27	6	24	5
	7	188	1	178	0
	% Error	*4.2 %*	*10.26 %*	*4.44 %*	*9.4 %*

The single tree results highlight the similarities between binary substitution and the Gower dissimilarity, as BS-MRT and Gower Db-MRT produced the same terminal nodes but with an exactly opposite tree and show marked similarities in the MDS locations plots (Figure 65, Figure 67). The consensus based methods show the same first split using variable "Cell.size" as the single tree methods, however binary substituted RF shows a different decision point to the treeboost (Figure 70, Figure 74).

Interestingly both consensus based methods find more complex trees, however only in the case of binary substituted treeboost does this translate into improved performance. Surprisingly binary substituted RF performs worst of all other methods (Table 24). By observation of the misclassification tables it is clear that binary substituted RF is strongly biased towards the malignant group in the training set, to the detriment of

overall predictive performance. The best performing model of the base methods is clearly treeboost, with the lowest classification error and a clear proximity image.

7.2.6 Global MCT

Global MCT random forests are grown on each variable in the training set with the following parameters:

(a) A separate random forest test set of 70 training set observations is removed before the analysis to tune the model.

(b) The bootstrapped sample that is used to grow each tree consists of 196 observations and 3 variables.

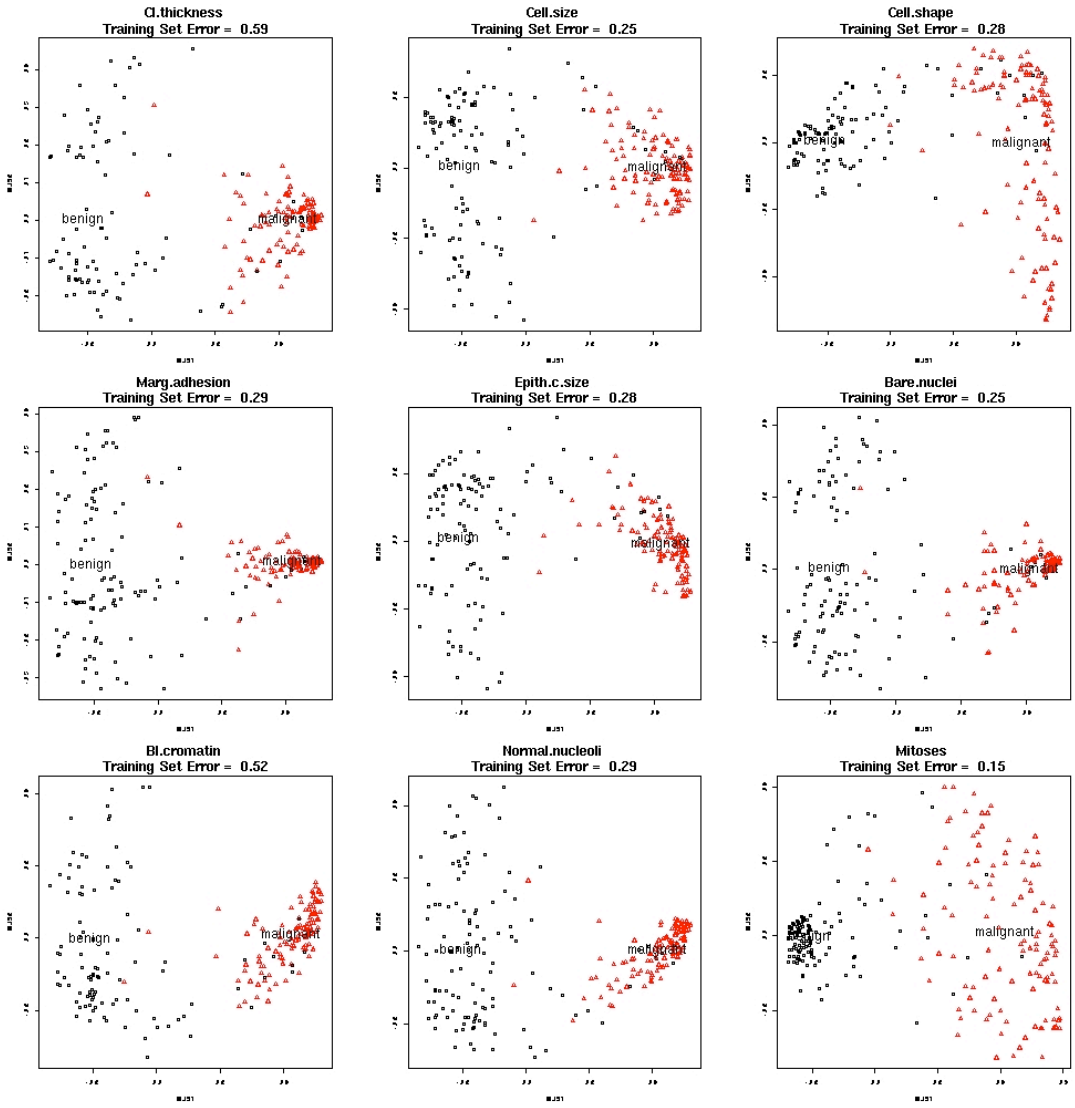
(c) A maximum tree size of 10 splits within the forest is allowed.

(d) The minimum terminal node size for each tree within the forest is 10 observations.

(e) There are 200 trees within the random forests.

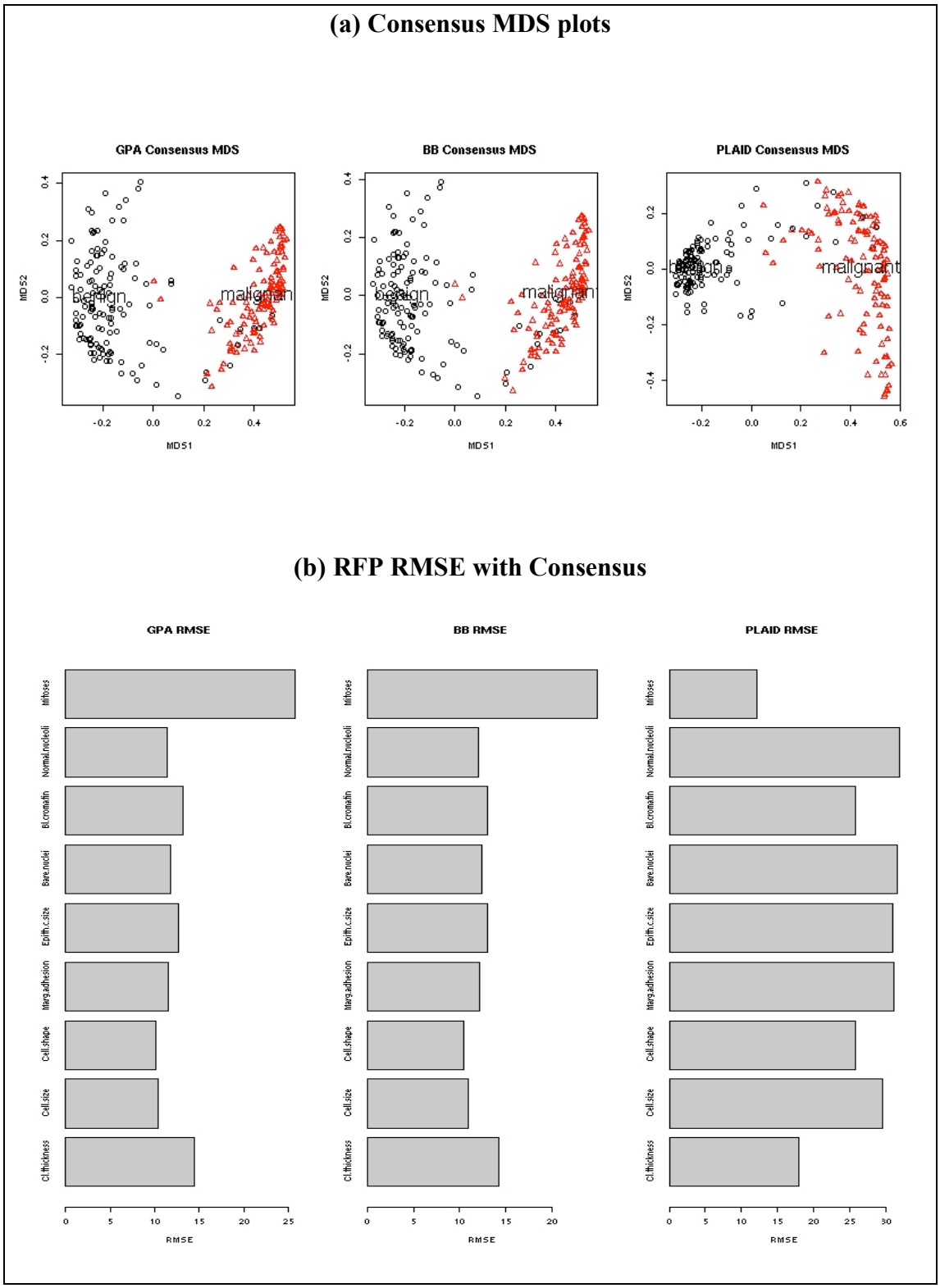
The individual RFPs (Figure 76) clearly show that the forests are finding a clear distinction between benign and malignant cancer groups. The training set performance for predicting each variable by the forest as a misclassification error is printed in the plot titles.

Figure 76: Breast cancer analysis individual RFP MDS plots.



Each RFP combination method is used to construct a consensus matrix (Figure 77(a)). By observation of the MDS plots, it appears that BB and GPA find similar configurations, and PLAID finds a different structure. This observation is reinforced by RMSE plots between the individual RFPs and the consensus (Figure 77(b)). From the RMSEs it can be seen that BB and GPA clearly favour the middle variables, performing poorest on Mitoses and Cl.thinckness, whereas PLAID favours these variables at the expense of the others.

Figure 77: Breast cancer analysis consensus MDS plots and RMSEs.

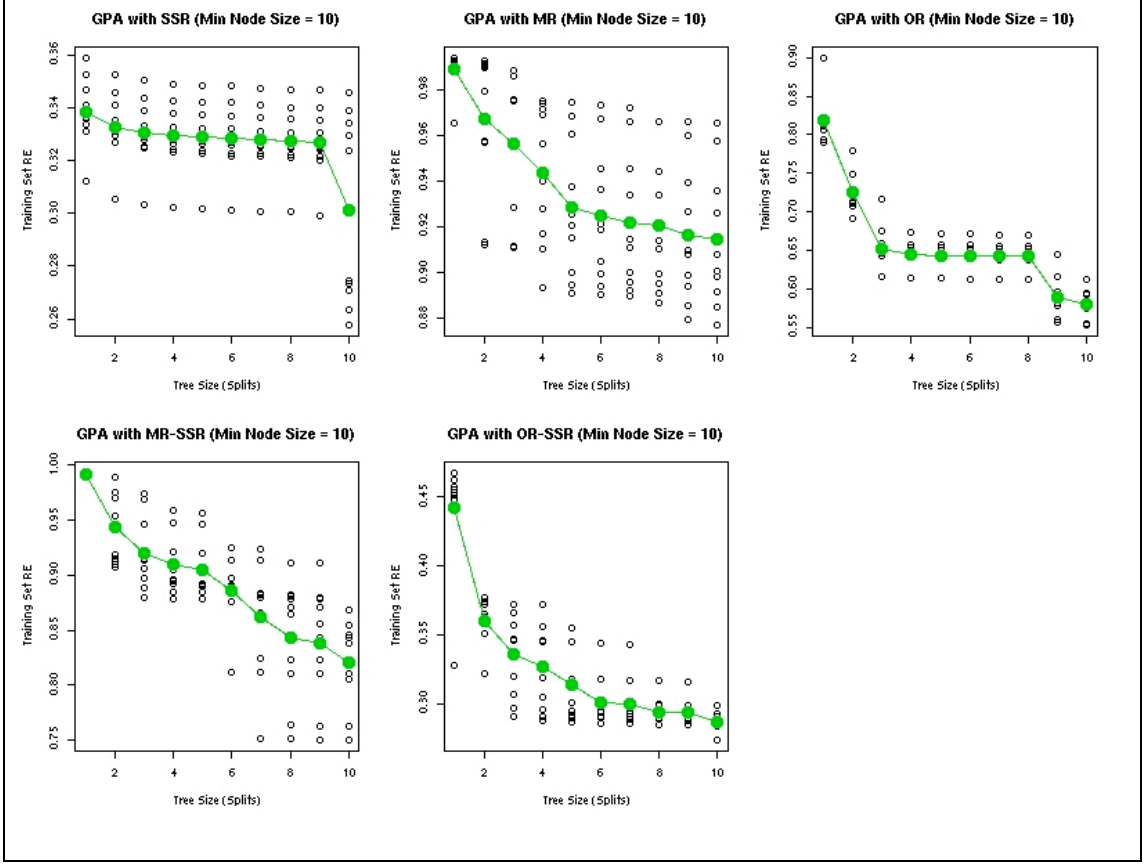


To assess how far to grow the MCT 10-fold CV is performed on the consensus matrix, with a minimum terminal node size fixed to 10 observations (Figure 78).

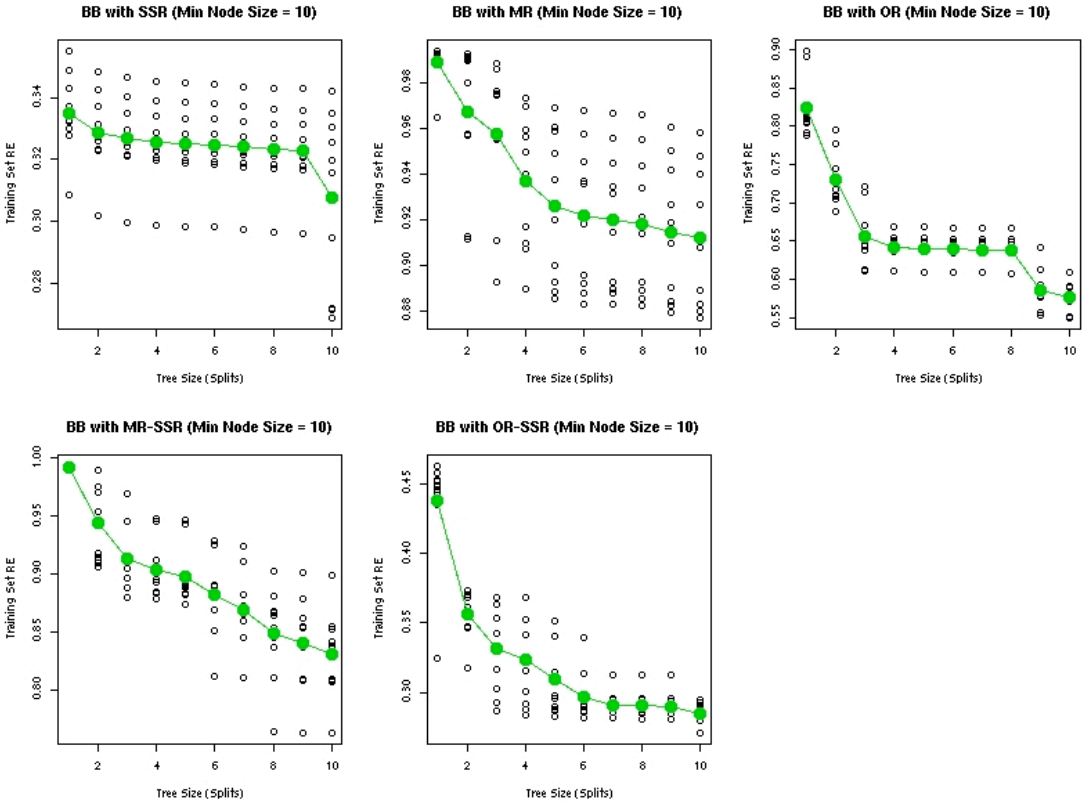
Clearly the best structure is resolved by the splitting method OR-SSR, which finds a RE elbow for GPA (Figure 78a) and BB (Figure 78b) at 6 splits, and for PLAID (Figure 78c) at 5 splits, with a corresponding mean RE of between 0.2 and 0.25. This RE equates to an R^2 of approximately 0.7, meaning the predicted consensus matrix accounts for between 65 % and 75 % of the consensus variation.

Figure 78: Breast cancer analysis global MCT 10-fold CV RE curves.

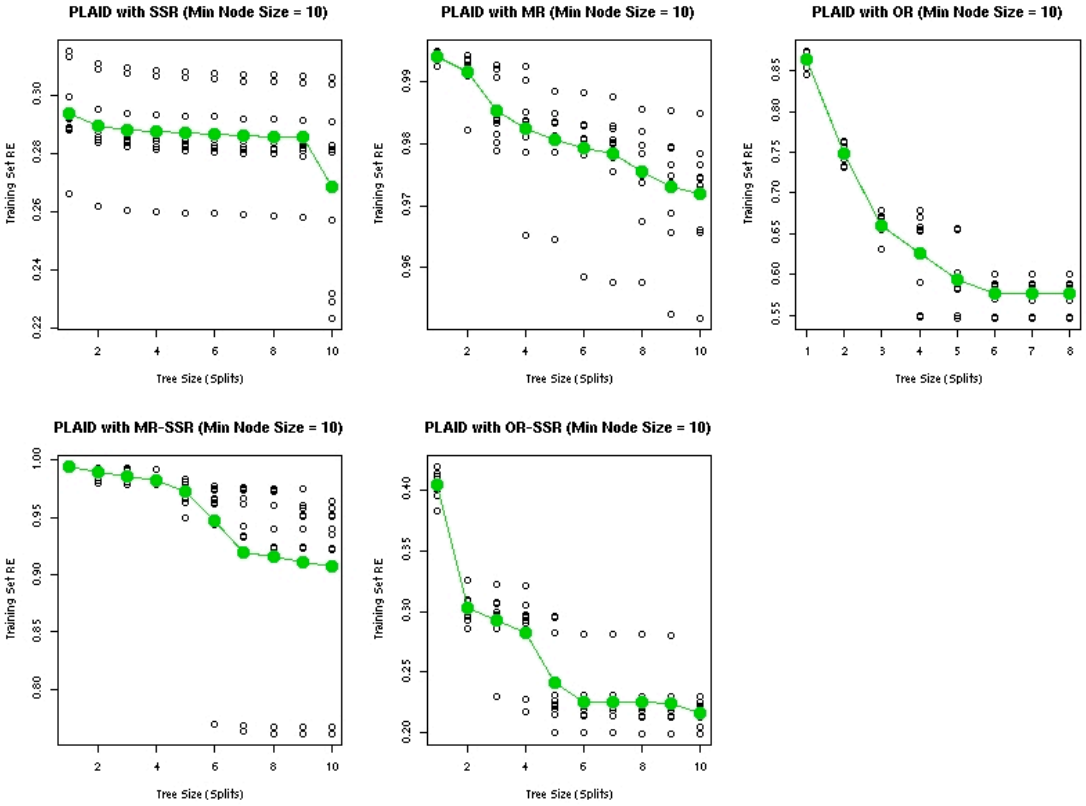
(a) GPA



(b) BB


BB with SSR (Min Node Size = 10)
BB with MR (Min Node Size = 10)
BB with OR (Min Node Size = 10)
BB with MR-SSR (Min Node Size = 10)
BB with OR-SSR (Min Node Size = 10)
Training Set RE
Tree Size (Splits)

(c) PLAID

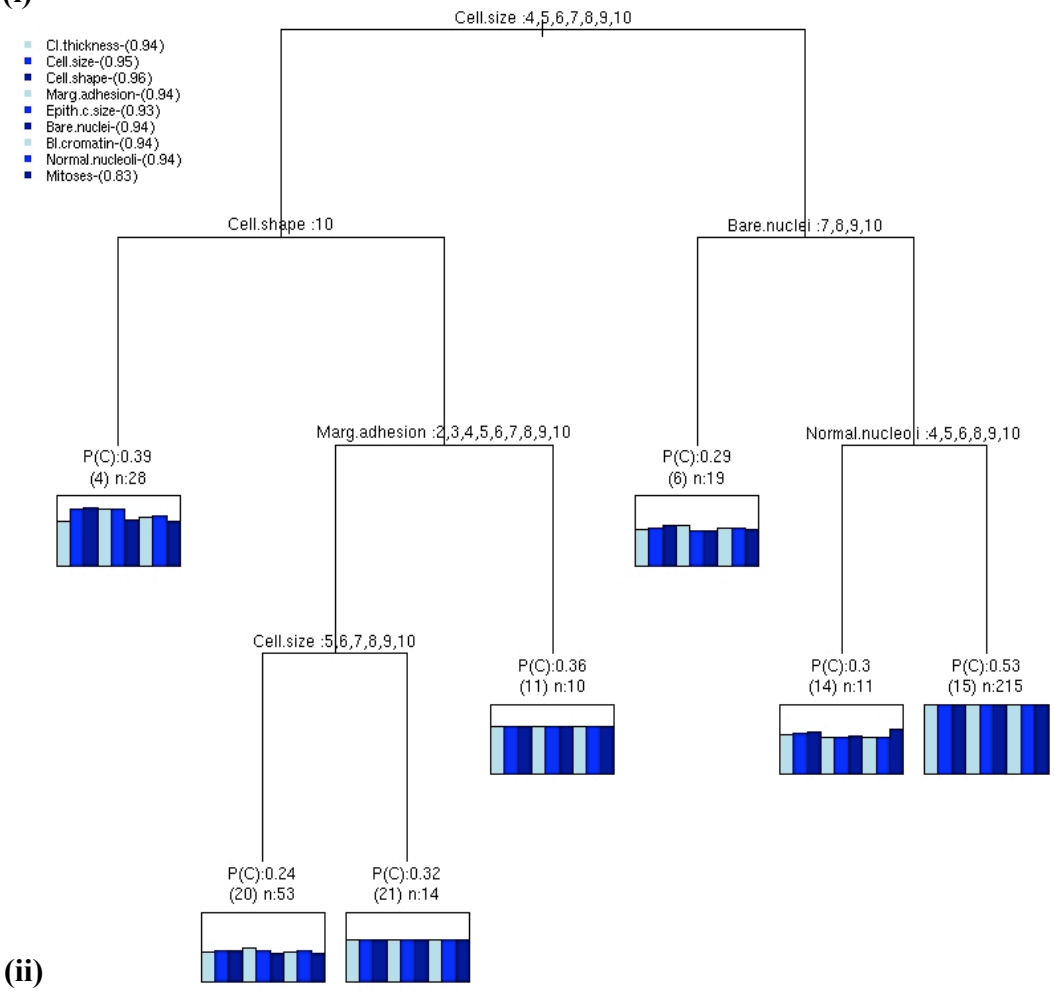

PLAID with SSR (Min Node Size = 10)
PLAID with MR (Min Node Size = 10)
PLAID with OR (Min Node Size = 10)
PLAID with MR-SSR (Min Node Size = 10)
PLAID with OR-SSR (Min Node Size = 10)
Training Set RE
Tree Size (Splits)

The global MCTs for GPA and BB are grown to 6 splits (7 clusters) and PLAID to 5 splits (6 clusters) all using the OR-SSR splitting method. The trees for GPA and BB are the same (Figure 79(a,i)) and only show subtle differences from splits observed in the PLAID MCT (Figure 79(b,i)). The terminal node locations found by the trees are displayed on the MDS plots of the consensus matrices (Figure 79(a,ii) and Figure 79(b,ii)) for GPA, BB or PLAID respectively. From this it can be seen that most effort is spent identifying the malignant group, with the majority of the benign group being positioned in both trees in terminal node 15. In the left corner of the MCTs, the response variable importance list (YVIP) list can be found. The structure found in the YVIP matches the RMSE combination plots (Figure 77b). The mean of the consensus at each terminal node is printed below the terminal node as a probability, "P(C)", along with the terminal node number in brackets, and the number of training set observations within that node. A bar plot of the P(C)s of each individual RFP at each terminal node is also presented.

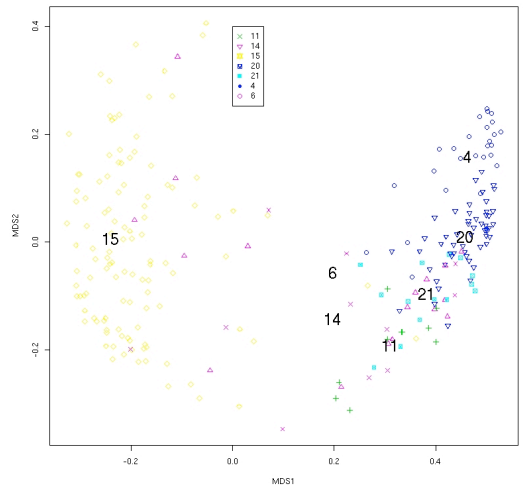
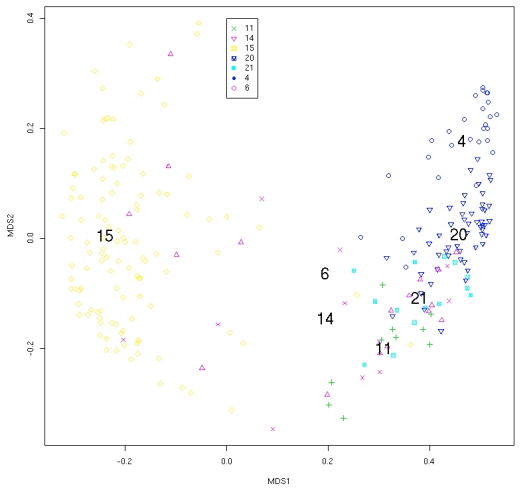
Figure 79: Breast cancer analysis best global MCTs and terminal node location MDS plots.

(a) GPA & BB OR-SSR MCT Tree

(i)

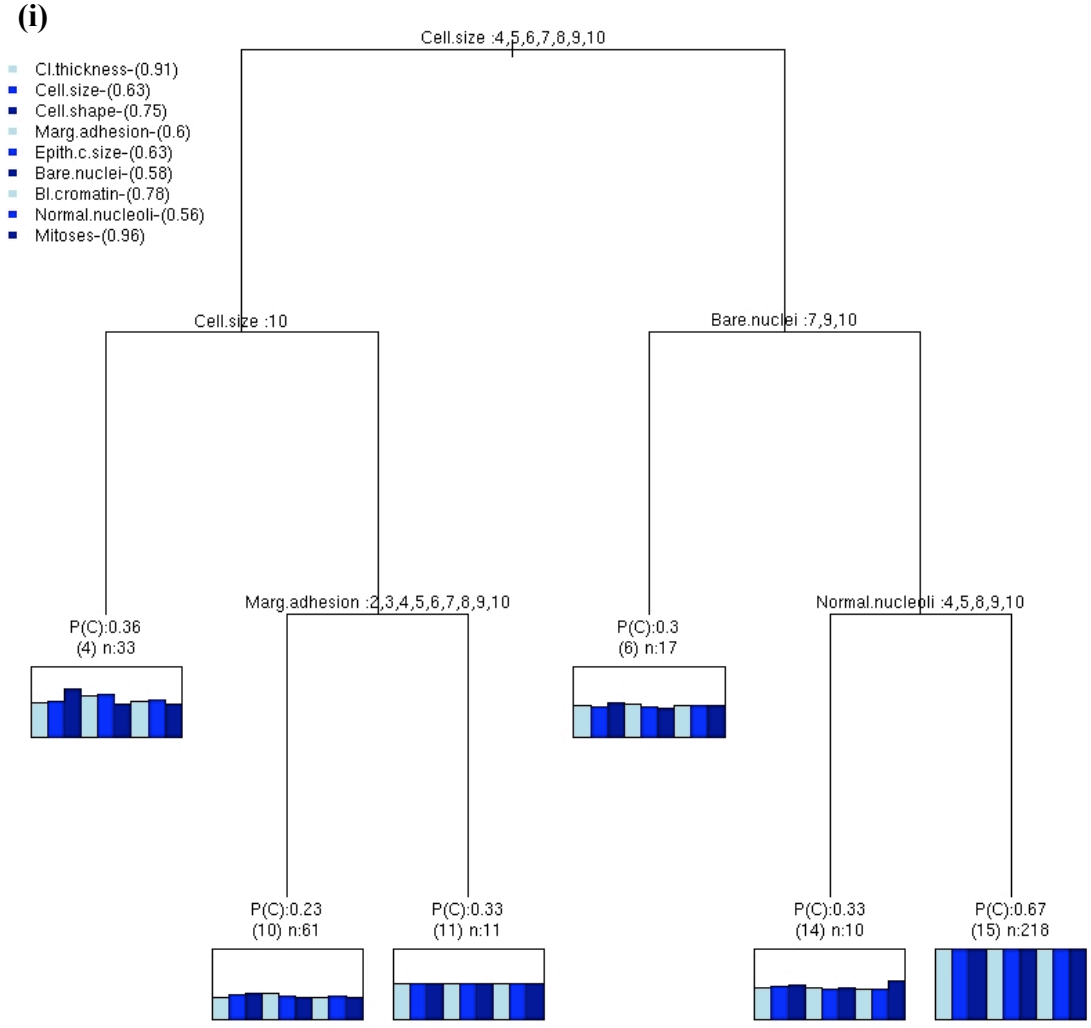


(ii)

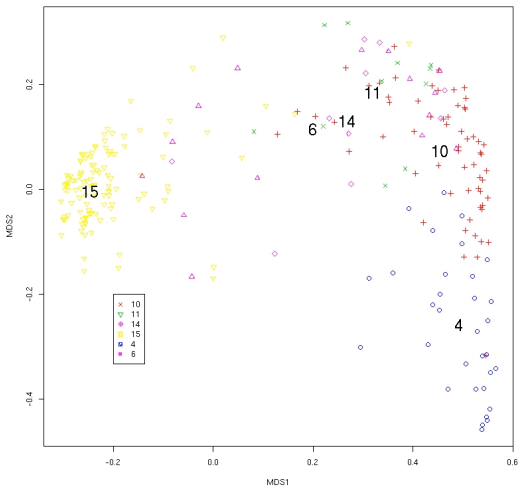
GPA MCT Terminal Node Locations	BB MCT Terminal Node Locations
	

(b) PLAID OR-SSR MCT Tree

(i)



(ii) **Plaid MCT Terminal Node Locations**



The training and test set performances for global MCTs (Table 25) show an overall test sample misclassification rate of approximately 4.9 %. When compared to supervised classification on the same data, a decision tree grown to 4 splits gives test set misclassification rate of 6.3 % and a random forest gives a test set misclassification rate of 2.8 %. Therefore the performance of global MCTs for the breast cancer dataset is approaching that of a random forest.

Table 25: Breast cancer analysis global MCT misclassification performances.

(a) GPA & BB (6 % Overall Error)

MCT Node	**Train Set (6.3 % misclassification)**		**Test Set (4.9 % misclassification)**	
	Benign	**Malignant**	**Benign**	**Malignant**
4	0	28	0	30
6	6	13	7	8
11	3	7	2	5
14	5	6	0	3
15	213	2	213	5
20	2	51	1	60
21	4	10	2	13
Overall Misclassification	*8.6 %*	*1.7 %*	*5.33 %*	*4 %*

(b) PLAID (5.4 % Overall Error)

MCT Node	**Train Set (6 % misclassification)**		**Test Set (4.87 % misclassification)**	
	Benign	**Malignant**	**Benign**	**Malignant**
4	0	33	0	34
6	6	11	7	8
10	6	55	3	70
11	3	8	2	4
14	3	7	0	3
15	215	3	213	5
Overall Misclassification	*8.0 %*	*2.6 %*	*5.33 %*	*4 %*

7.2.7 Local MCT

As local MCTs build a separate forest at each node, the random forest parameters are presented in percentages. The local MCT parameters are set at the following:

(a) The bootstrapped sample used to grow each tree within each node is defined as 70% of node observations and 33% of variables.

(b) Maximum random forest tree size is 3 splits.

(c) Minimum MCT and random forest terminal node size is 10 observations.

(d) Random forest size is 200 trees.

(e) OR-SSR splitting criteria.

RE and AIC are generated to assess local MCT tree size. For a fair comparison with global MCTs only OR-SSR splitting criteria is employed as it clearly outperformed other splitting criteria in global MCTs for this problem. Local MCTs are run using all three RFP combination methods.

For local MCTs the results for each combination method with OR-SSR splitting are identical. The RE and AIC plots each indicate a tree size of 3 split or 4 groups (Figure 80) and the resulting MCT tree for each combination method at 3 splits is the same (Figure 81i). This results in the same misclassification performance of 5.44 % error on the test set (Table 26). The only difference in the trees is the subtle differences observed in the PLAID MDS plot of the ACM matrix when compared to either the BB or GPA plots (Figure 81ii).

Figure 80: Breast cancer analysis local MCT RE and AIC plots.

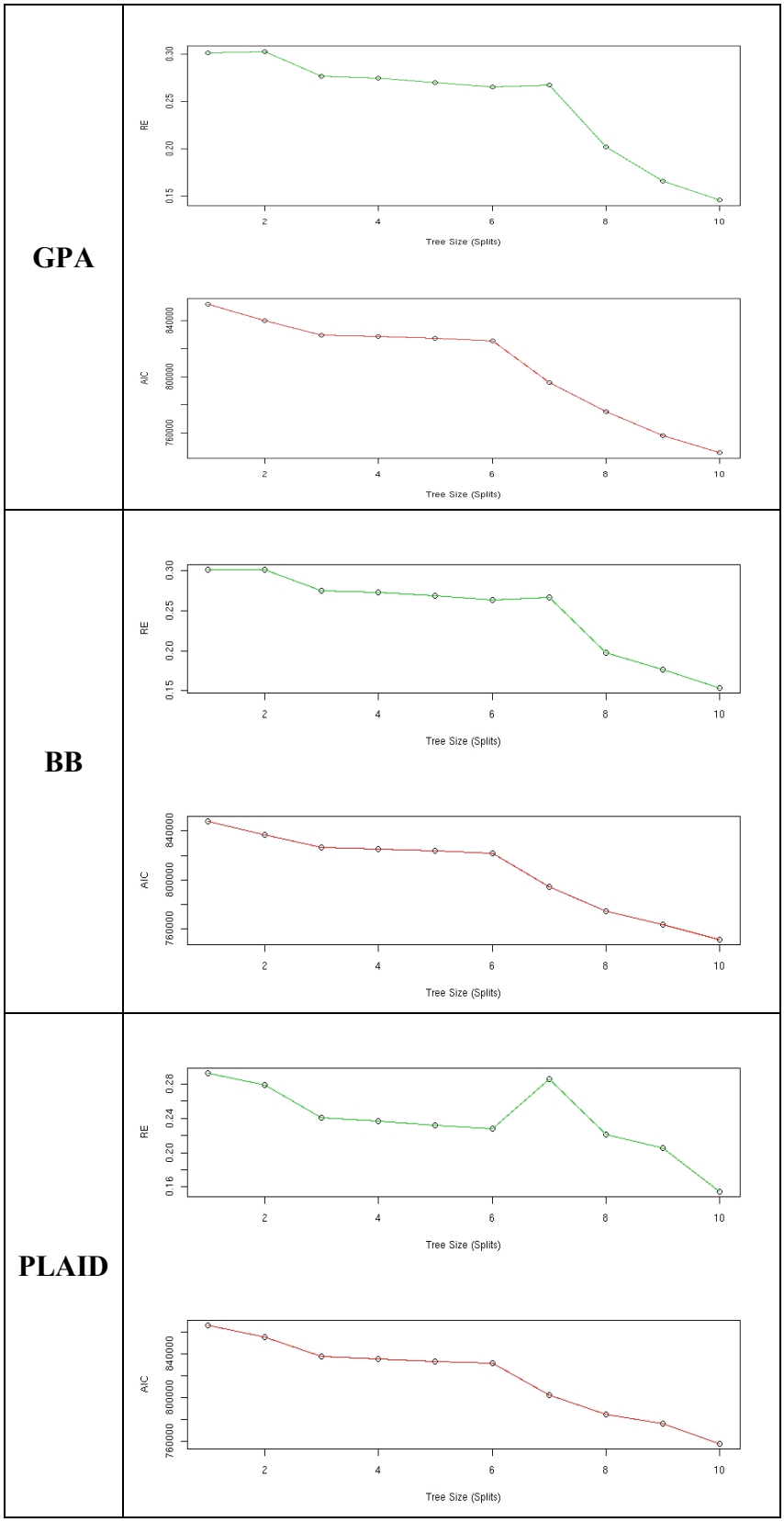
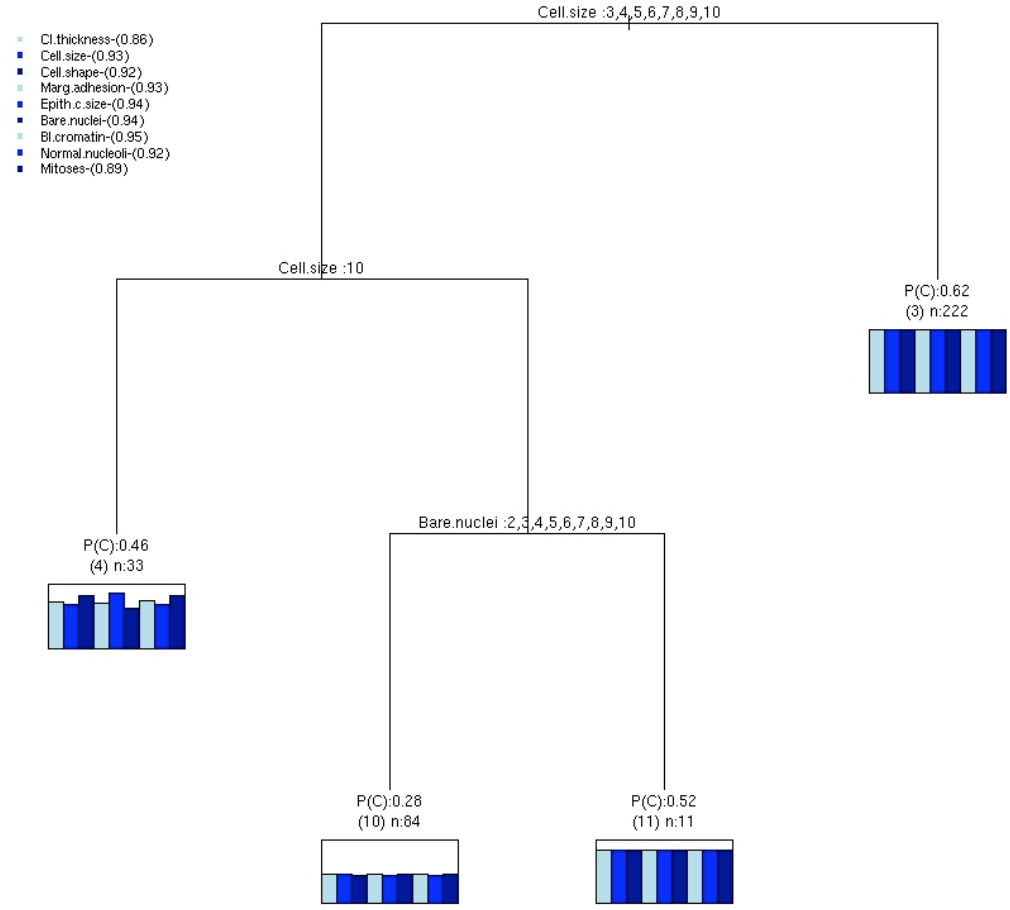


Figure 81: Breast cancer analysis GPA, BB and PLAID, OR-SSR local MCT.

(i) GPA, BB and PLAID local MCT



(ii) MDS Terminal Node Location Plot

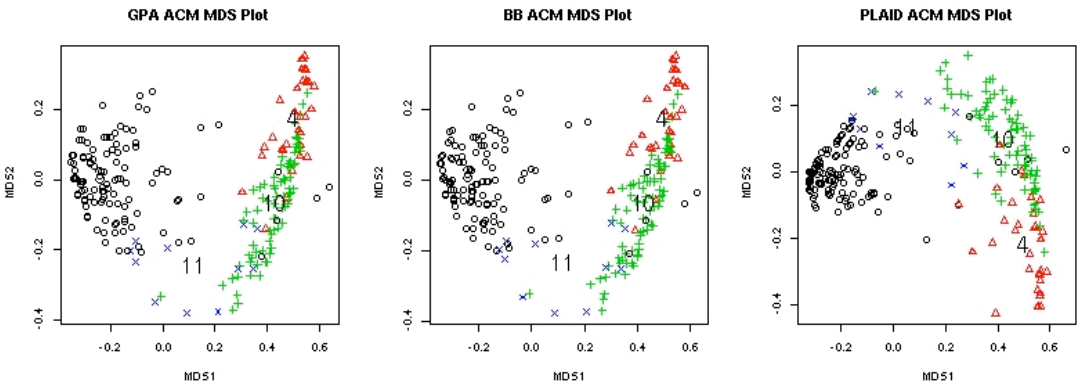


Table 26: Breast cancer analysis local MCT misclassification performances.

MCT Node (5.57 % Total Error)	Train Set (5.71 % misclassification)		Test Set (5.44 % misclassification)	
	Benign	**Malignant**	**Benign**	**Malignant**
3	215	7	202	5
4	0	33	0	34
10	10	74	10	81
11	8	3	13	4
Overall Misclassification	*4.29 %*	*8.54 %*	*4.44 %*	*7.26 %*

7.2.8 Breast cancer summary

Of all the methods presented, global MCTs using the PLAID consensus produced the most accurate tree (test set misclassification performance 4.87 %) (Table 25). Furthermore, the performances of all MCT methods are better than any base tree method. Compared to existing literature on this dataset MCTs are performing comparably. Clustering using SOM achieved a misclassification rate of 4.68 % (Pantazi, Kagolovsky and Moehr 2002) however this method assumes all variables are ordinal and provides no measures of variable importance. Supervised analysis of this dataset has been shown to perform well below 10 % misclassification, with a linear programming approach achieving 3 % misclassification (Mangasrian and Wolberg 1990).

Given that there are two groups (benign and malignant) within the dataset, the most accurate models in this case were Gower Db-MRT and binary substituted MRT as they found 2 terminal nodes. As accuracy increased so did tree size with the most accurate MCT identifying 6 groups within the data. This inflation of group number is due to the overlap between the two groups. This is reinforced by observation of the consensus MDS images, as in all plots expect AA-Treeboost two groups are obvious.

These results imply that the simple base methods (Gower Db-MRT and binary substituted MRT) do not have sufficient power to identify the overlapping groups.

The improvement in resolution gained from a local MCT should also be noted. All combination methods for local MCTs produced the same tree. Furthermore a smaller tree is obtained with comparable predictive performance. These results highlight the differences between local and global MCTs, and show that once tuned local MCTs produce a more accurate result.

7.3 Mixed Type Profiling: Horse Colic Dataset

In this analysis MCTs are used as a mixed type profiling tool. Here there is no known set of groups to compare against, and therefore the quality of the groups found must be assessed on how representative they are of each response variable. This analysis is performed on the horse colic dataset, where the response set comprises of variables that describe the observed physical state of each horse, and the predictor set are variables that describe the type, site and severity of their colic lesion (Mcleish and Cecile 1989). The goal is to use MCTs to identify groups in the predictor variables describing the lesions that correspond to groups within the response set of physical descriptors.

The horse colic dataset contains 300 observations on 17 variables, 5 being quantitative and 12 being either ordinal or nominal (Table 27). With such a complicated response set spanning many types, it is expected that some variables will display different group profiles. In this analysis MCTs are used as a search for subgroups of response variables that display a common group structure. To do this a recursive search for common group structure using plaid combining is described. The result of this search is subgroups of response variables that have similar configurations within their RFPs. Upon these subgroups, separate MCTs are grown and compared to the structure found in an overall MCT involving all RFPs. This is a data reduction step that is aimed at improving the understanding of group structure within each variable and how it relates to the overall structure within the dataset.

Table 27: Horse colic analysis dataset description.

Variable Set	Variable Name	Description	Type	Missing Values
Response	REC.TEMP	*Rectal temperature*	Continuous	60
Response	PULSE	*Pulse rate*	Continuous	24
Response	CELL.VOL	*Packed cell volume*	Continuous	29
Response	TOT.PROT	*Total protein*	Continuous	33
Response	RESP.RATE	*Respiratory rate*	Continuous	58
Response	TEMP.EXT	*Temperature of extremities*	Ordinal {4 levels}	56
Response	PERIF.PU	*Peripheral pulse*	Ordinal {4 levels}	69
Response	MUCOUS.M	*Mucous membranes*	Nominal {6 levels}	47
Response	CAPILL.R	*Capillary refill time*	Ordinal {2 levels}	34
Response	PAIN	*A subjective judgment of pain level*	Nominal {5 levels}	55
Response	PERISTAL	*Peristalsis*	Nominal {4 levels}	44
Response	ABDOM.DI	*Abdominal distension*	Ordinal {4 levels}	56
Response	NASO.REF	*Nasogastric reflux*	Ordinal {4 levels}	106
Predictor	LESION	*Is surgery required on the lesion*	Dichotomous Yes or No	0
Predictor	LESION.S	*Site of the lesion*	Nominal 1. Gastric 2. Small intestine 3. Large colon 4. Large colon and cecum 5. Cecum 6. Transverse colon. 7. Retum/descending colon 8. Uterus 9. Bladder 10. All intestinal sites 11. None	0
Predictor	LESION.T	*Type of the lesion*	Nominal 1. Simple 2. Strangulation 3. Inflammation 4. Other	60
Predictor	LESION.A	*Subtype of the lesion*	Nominal 1. Mechanical 2. Paralytic 3. N/A	1

7.3.1 MRT methods

To begin analysis on the horse colic dataset, simple MRTs with the Gower distance matrix and binary substituted response sets are grown. If the grouping structure within the response is strong then these methods will adequately describe the groups present. However it is expected that with such a complicated response these methods will be insufficient and unable to find meaningful structure.

7.3.1.1 Gower dissimilarity Db-MRT

Figure 82: Horse colic analysis Gower Db-MRT RE curve.

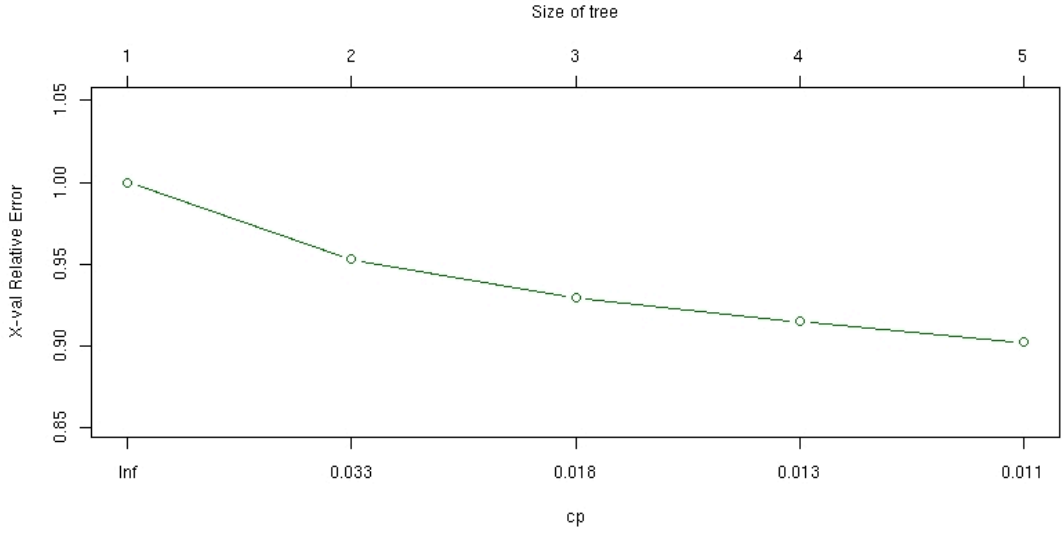
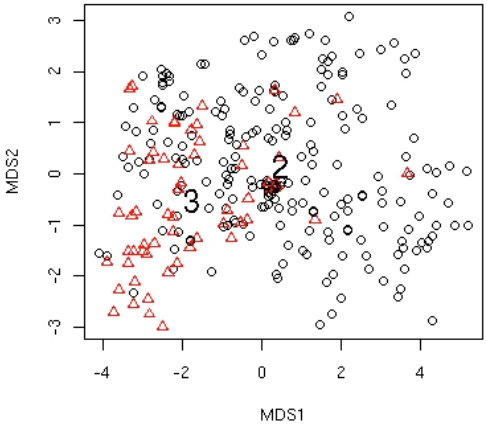


Figure 83: Horse colic analysis Gower Db-MRT and terminal node locations.

Gower Db-MRT	MDS plot of the Gower distance matrix.
	

7.3.1.2 Binary substituted MRT

Figure 84: Horse colic analysis binary substituted MRT RE curve.

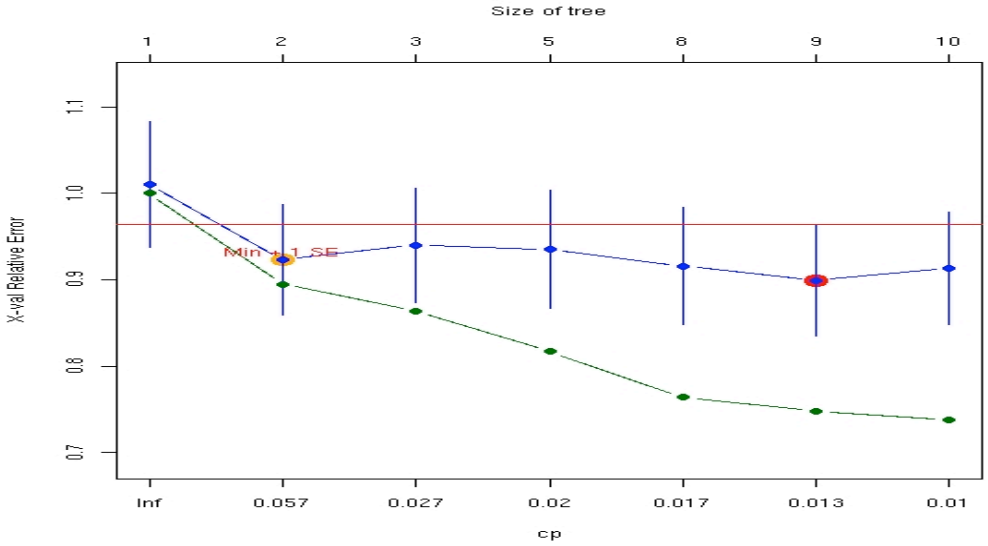
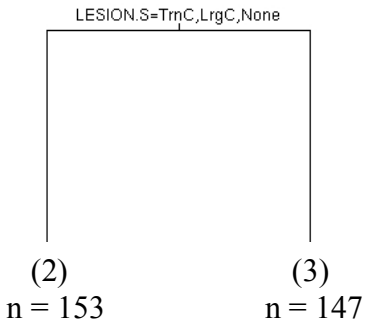
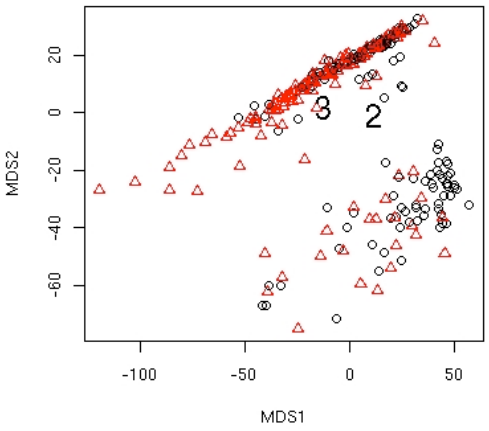


Figure 85: Horse colic analysis binary substituted MRT and terminal node locations.

Binary Substituted MRT	MDS plot of the binary substituted response using a Euclidean distance.
	

7.3.1.3 MRT method summary

The issue of missing values within the response set is primary when interpreting the MRT methods. The Gower distance simply ignores comparisons that involve a missing value in its distance computation. The result of such an approach is no observable grouping structure within the response set (Figure 83). This lack of structure is represented by a high RE of 95 % (Figure 82) and results in a simple single split tree (Figure 83).

For the binary substituted data, the missing values are imputed using a K-nearest neighbour averaging on a Euclidean distance (Hastie, Tibshirani, Narasimhan and Chu 2005). The effect of this is a more obvious grouping structure within the MDS plot, which is not found by the MRT (Figure 85). The MRT itself acknowledges this with a poor predictive performance, displaying a RE of 0.93 +/- 0.052 (Figure 84).

The implications of these results are that the groups within the profiling set are not obvious within the predictor set. As a result the trees found are simple and poorly performing.

7.3.2 Tree-based ensemble methods

To benchmark the MCT methods, overall consensus matrices are produced using random forests and treeboost on the binary substituted response. The important results of these techniques will be observable structure within the MDS plots of the ensemble proximity matrices, and measures of predictive accuracy and stability of tree based methods with the error convergence plots.

The consensus approaches show a much improved resolution of the lesion groups within the response (Figure 89, Figure 93). However the complexity of these relationships is highlighted with the random forest models requiring over 200 trees to become stable and treeboost over 100 (Figure 86, Figure 90). The partitions of the proximity images show for both methods a clear 2-3 group structure (Figure 87, Figure 91).

The MCTs for each ensemble proximity matrix are slightly different (Figure 88, Figure 92) with nodes 6 and 7 being found by lesion type in random forest ensemble MCT splitting and by whether the lesion was surgical or not, in treeboost ensemble MCT. For the split, of the 106 strangulation lesion types in LESION.T, 98 of these are flagged as being surgical in LESION implying a strong overlap between the two

potential splits. The total difference between the two splits is 28 observations, which is 9.34 % of the observations.

7.3.2.1 Binary substituted random forests

Figure 86: Horse colic analysis binary substituted RF error convergence plot.

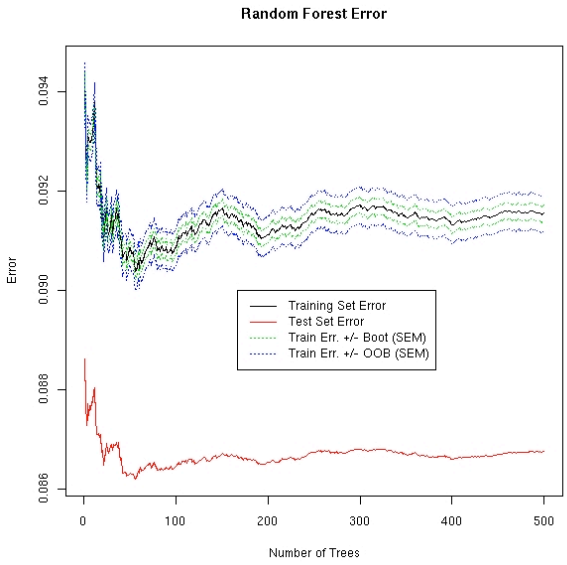


Figure 87: Horse colic analysis binary substituted RF RE curves.

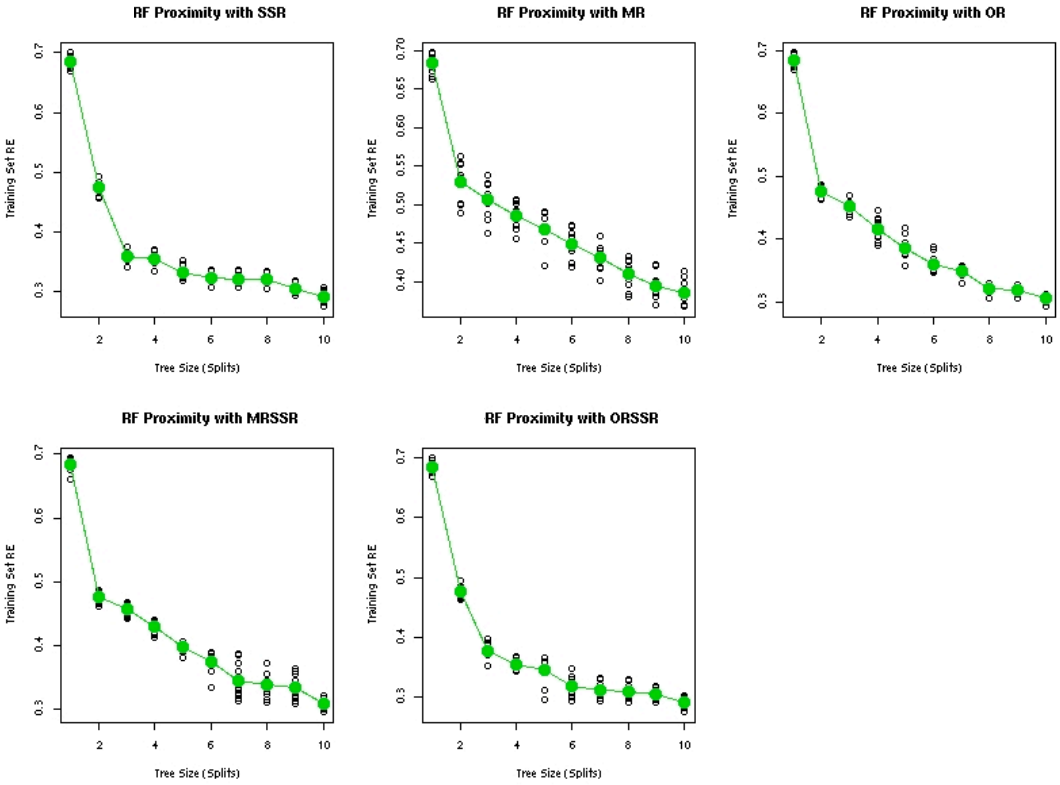


Figure 88: Horse colic analysis RF tree grown to 3 splits using SSR splitting.

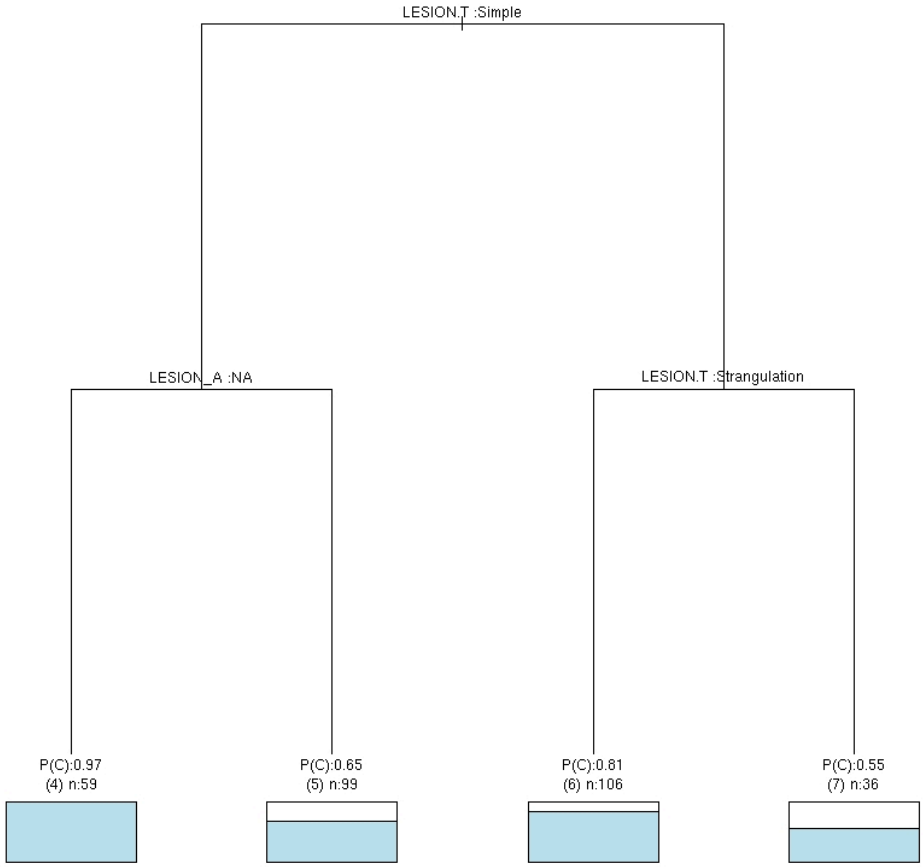
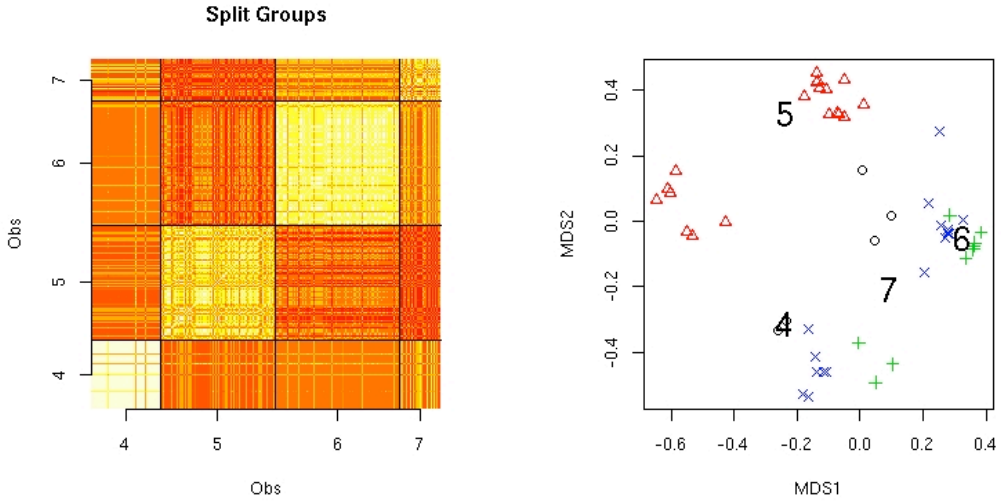


Figure 89: Horse colic analysis RF proximity images.



7.3.2.2 Binary substituted treeboost

Figure 90: Horse colic analysis binary substituted treeboost error convergence plot.

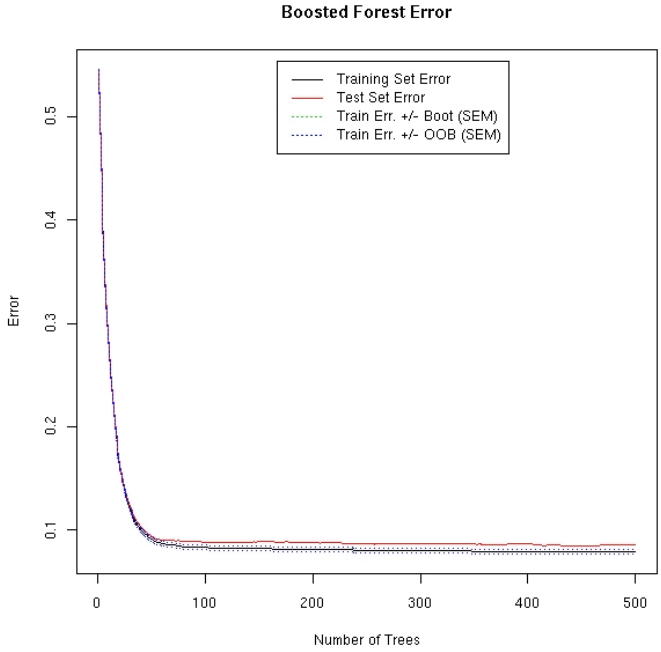


Figure 91: Horse colic analysis binary substituted treeboost RE curves.

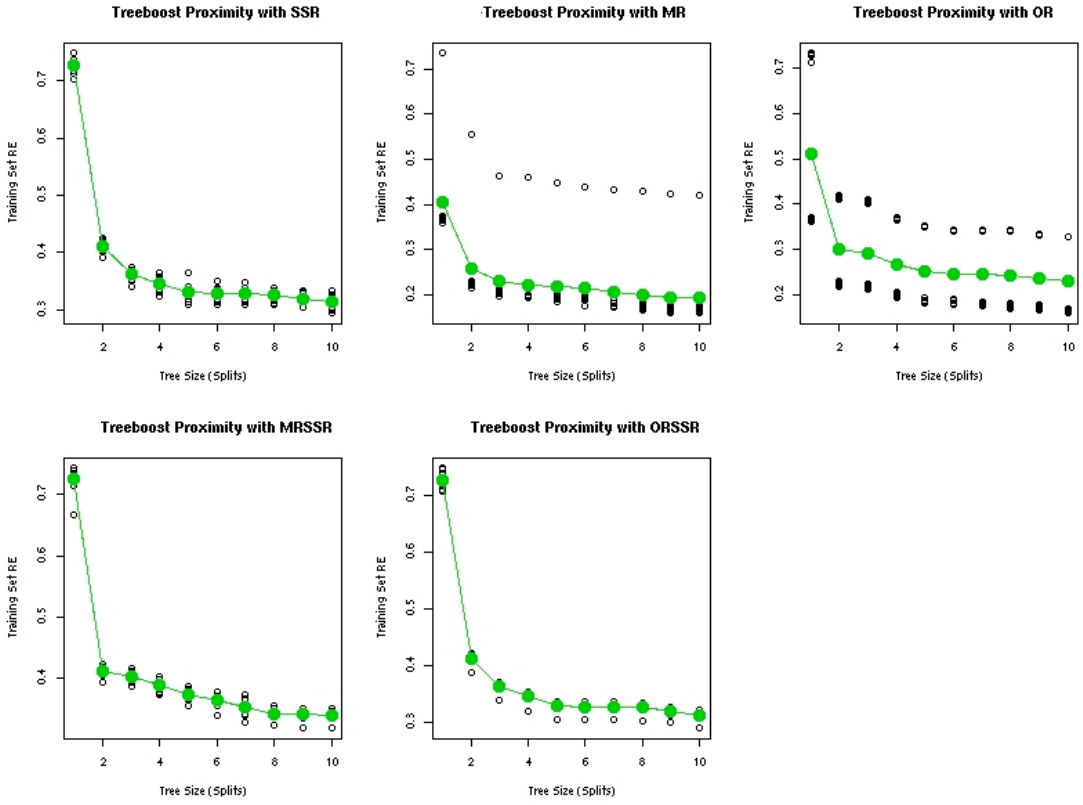


Figure 92: Horse colic analysis treeboost tree grown to 3 splits using SSR splitting.

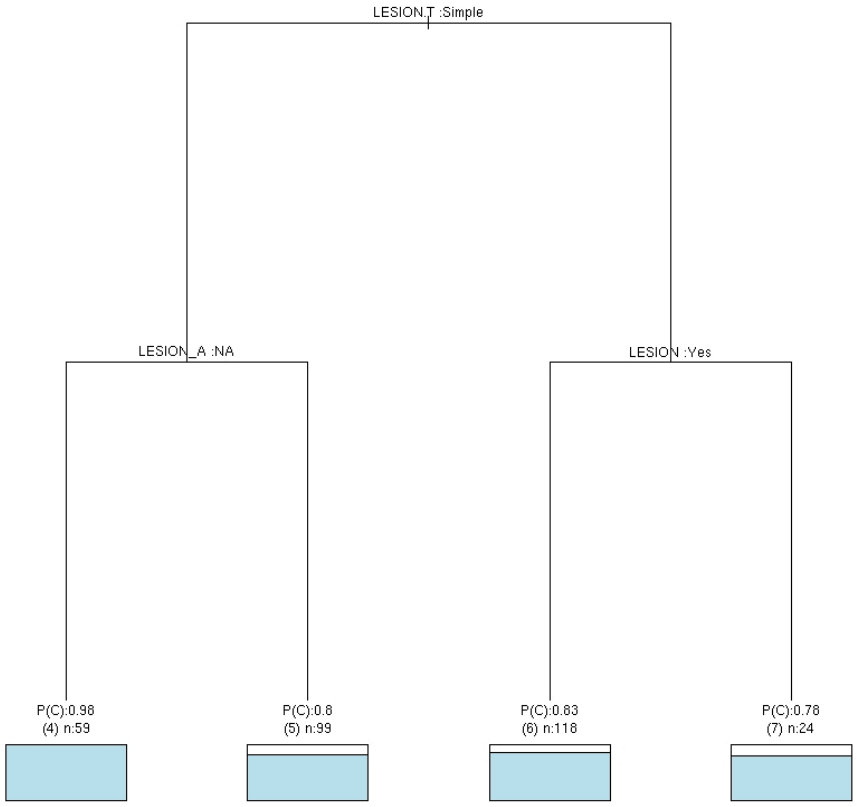
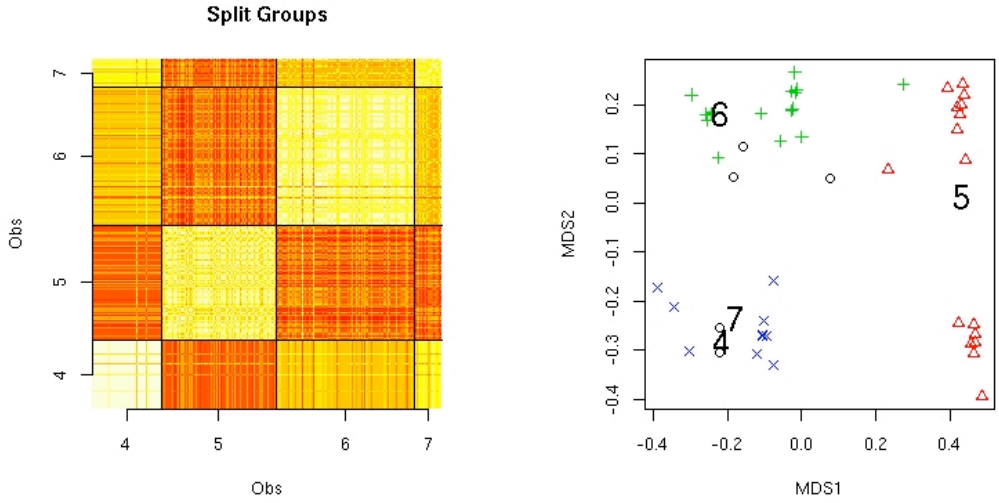


Figure 93: Horse colic analysis treeboost proximity images.



7.3.3 MCT methods

MCTs treat each variable within the response set individually to gain more understanding on the grouping structure of each individual response. As a result MCTs can be used not only for finding common profiles that exist in the entire dataset, but also sub-profiles or groups that are present in only a subset of response variables. This analysis focuses on MCT's ability to find these sub-groups and improved understanding of the final groups gained through the filtering process.

The first step in this analysis is the construction of a global MCT to profile the complete response set of the horse colic dataset. On this terminal node filtering is performed. This will show that not all response set variables express every node within the MCT. Secondly an algorithm of filtering the response variables before an MCT is grown is presented. This algorithm finds groups of variables within the profiling set using the PLAID consensus generation method. By doing this it is shown that further understanding and improved resolution of the groups found by the MCT is possible.

7.3.3.1 *Complete response set global MCT*

The first step in the analysis is to run the random forests for each response separately. These RFPs are used for all analysis. The global MCT random forest parameters:

- Set seed at 123.
- Separate test percentage of 60 observations to evaluate the ensemble's performance.
- 168 observations and 1 predictor used to construct each tree.
- Maximum tree size is 10 splits.
- Minimum terminal node size is 10 observations.
- The random forest is built to 200 trees.

Before being passed into any further analysis the performance of the random forests is assessed. The percent training set error in the title of the RFP images plot in Figure 94 show that for the response variables REC.TEMP, CELL.VOL, TOT.PROT and RESP.RAT the error in prediction is greater than if a simple mean is used as the prediction. As a result these variables are removed from the analysis.

To construct the global MCT the following profiling variables are used:

- PULSE
- TEMP.EXT
- PERIF.PU
- MUCOUS.M
- CAPILL.R
- PAIN
- PERISTAL
- ABDOM.DI
- NASO.REF

The consensus MDS plots (Figure 95a) show that the structure within the individual RFPs (Figure 94) has been maintained. Each combination method appears to have identified very similar structure with no observable difference in the MDS plots or in the RMSE profiles (Figure 95b). This similarity is unsurprising, as all individual RFP images appear to show similar profiles. The 10-fold global MCT RE graphs (Figure 96) indicate the best splitting function is SSR, and all show a full MCT size of 3 splits (4 groups is optimal). From this the MCT is grown with SSR to three splits using the GPA consensus (Figure 97).

Figure 94: Horse colic analysis individual RFP MDS plots. The MDS plots are coloured by the predictor variable LESION.

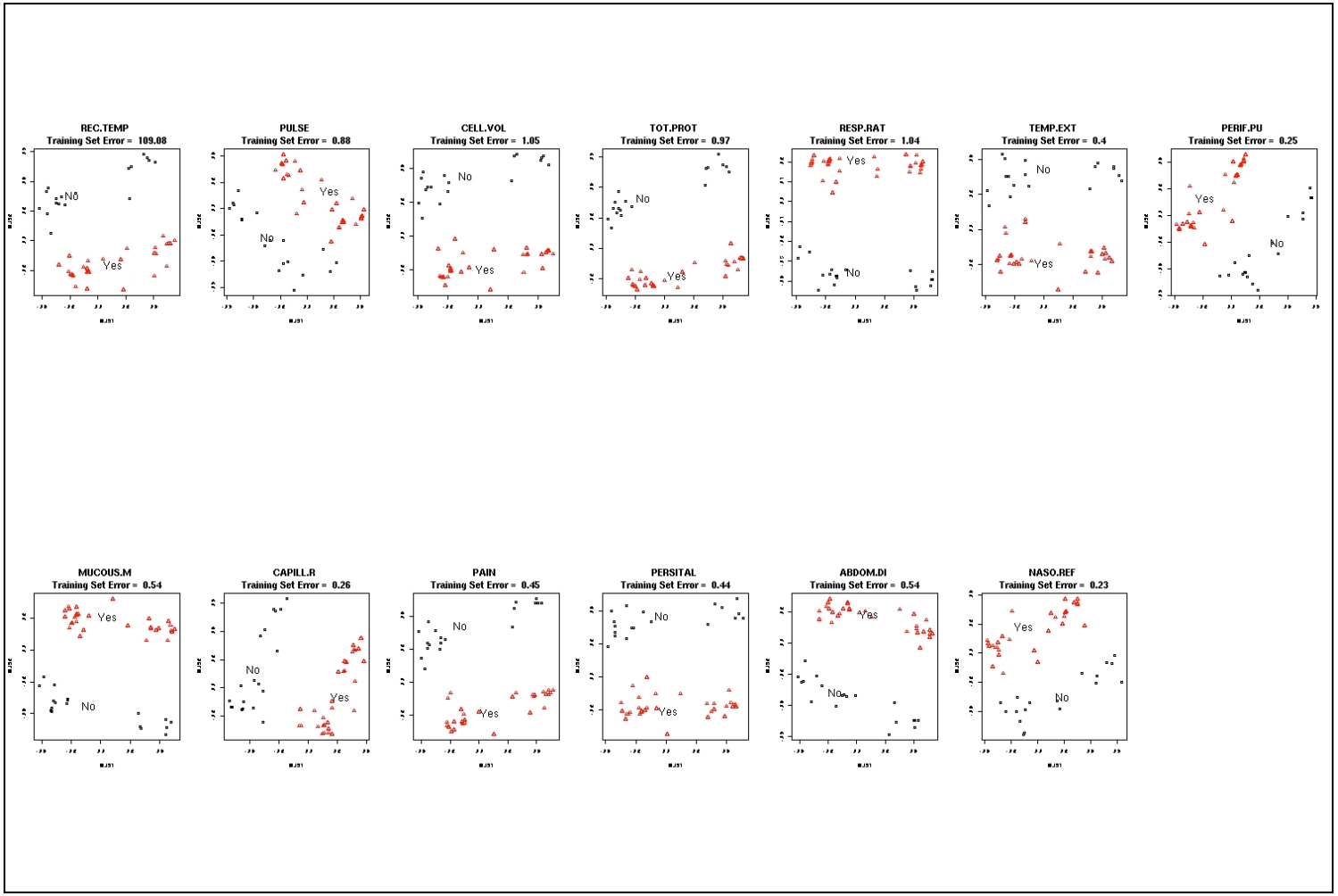


Figure 95: Horse colic analysis consensus MDS plots and consensus RMSE plots.

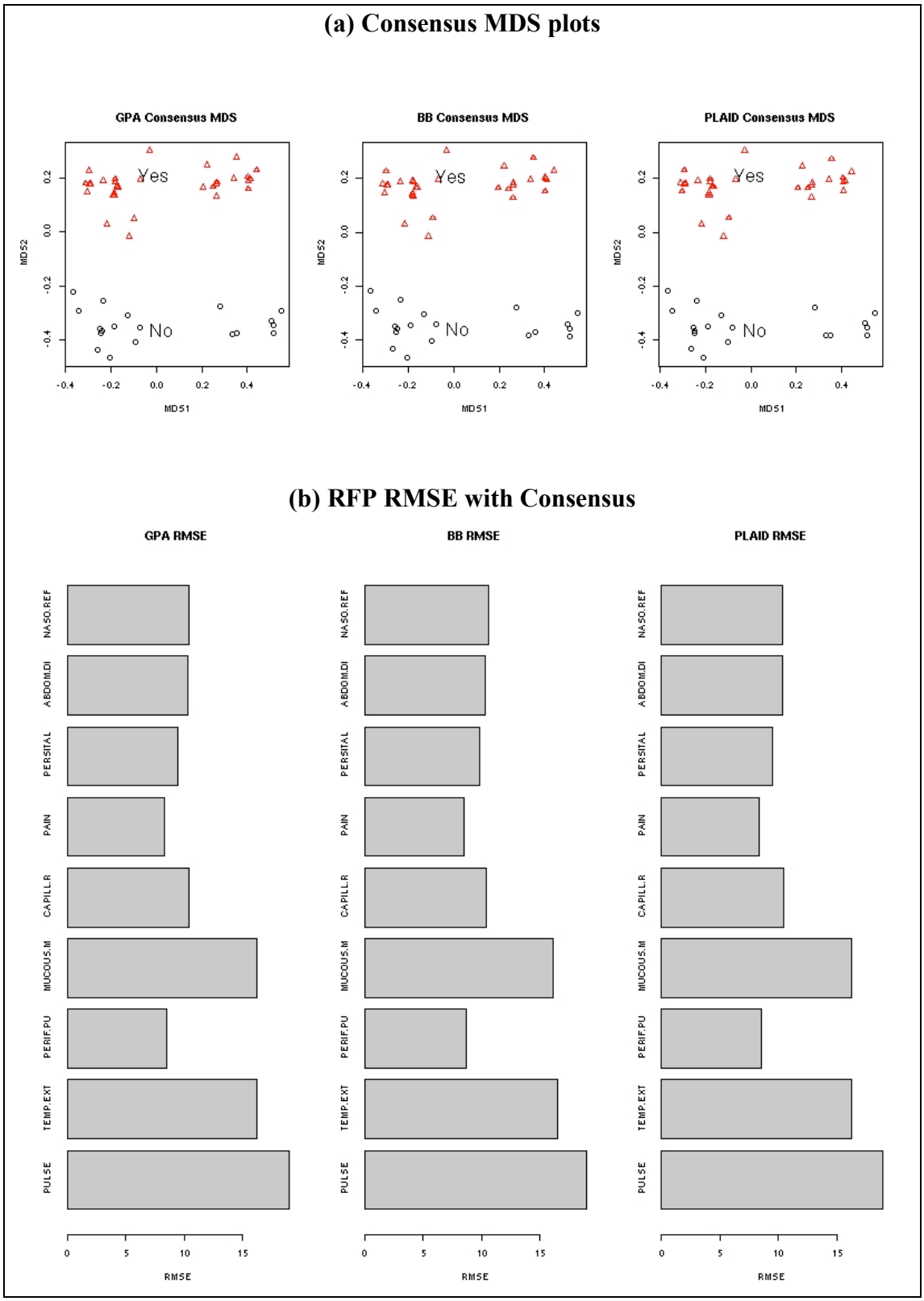
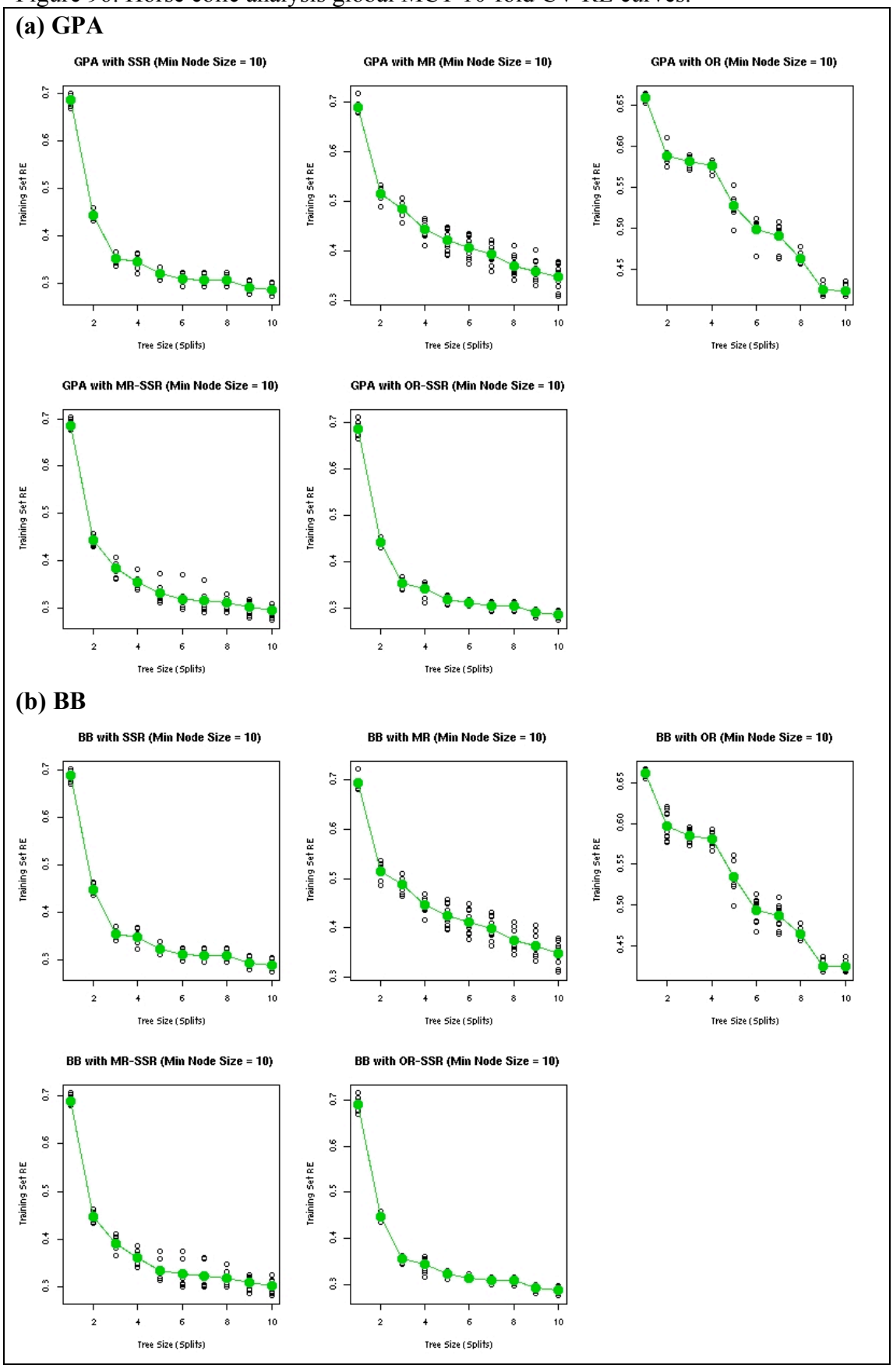
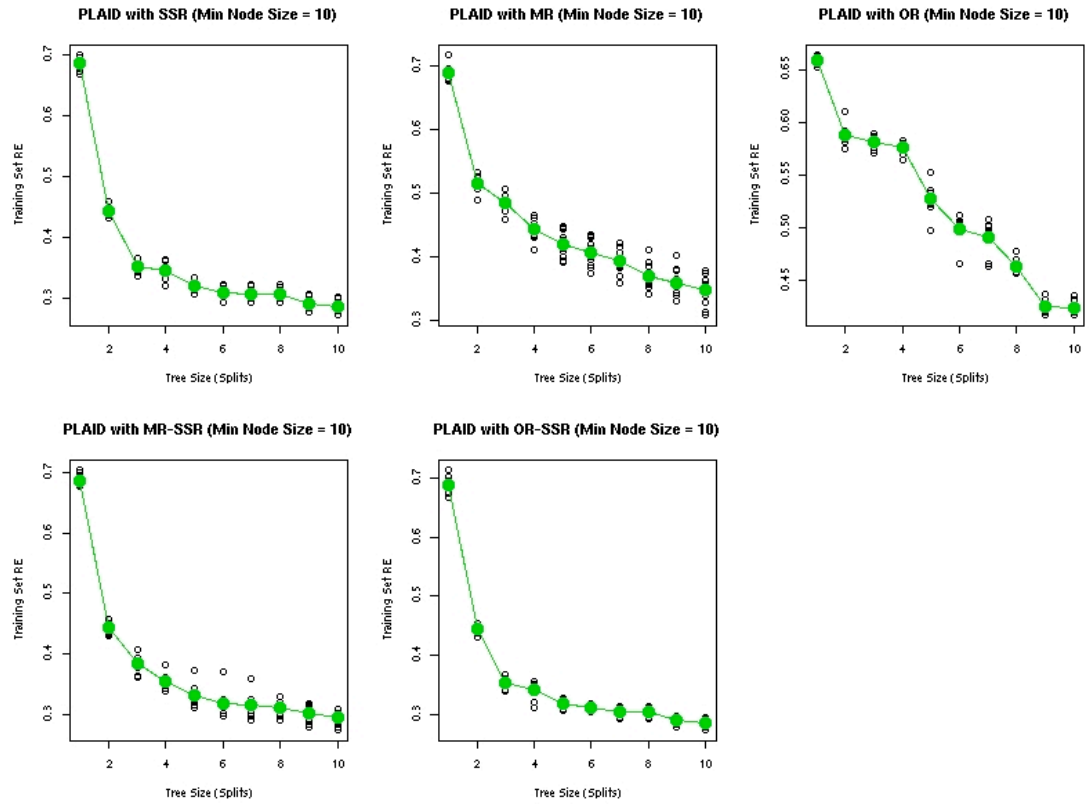


Figure 96: Horse colic analysis global MCT 10-fold CV RE curves.



(c) PLAID



Taking into consideration the similarity in consensus, individual RFP configurations and the RE curves, it is not surprising that for each combination method with SSR splitting grown 3 splits, the same MCT is produced (Figure 97a). Interestingly, the least obvious group in the MDS plot, (Figure 97b, group 4) is the most well expressed in the MCT, showing a within node probability of 0.97. Also, each group, especially group 5, appears to be a combination of two groups which have not been identified. In fact these groups can never be fully resolved, even when the MCT is grown to 10 splits shown in Figure 106.

7.3.3.2 Complete response set global MCT plaid terminal node filtering

Plaid terminal node filtering takes the sub-matrices for each terminal node for each RFP and observes their structure. At a terminal node it is assumed that each RFP displays the same structure. The assumption is that each cell can be modelled sufficiently with the mean centroid of that sub-matrix. The PLAID consensus generation is seen as a way to test for the validity of this assumption. If the plaid model finds a κ_m of '1', it means that this RFP has a different count profile to the other RFPs. If the same structure is found the plaid consensus is the mean of all consensus matrices and each κ_m will be zero.

Running plaid terminal node filtering upon an MCT gives an indication of which RFPs express each group. The result of this process (Table 28) identifies variables that express that node's consensus structure as '0'. For those that deviate, the magnitude and direction of the deviation is estimated. The results clearly show that terminal node 4 is the most stable node with only CAPILL.R and ABDOM.DI expressing different configurations. Conversely terminal node 5 is the least stable with only PULSE, PERIF.PU, PERSITAL and ABDOM.DI expressing the consensus structure. Interestingly, terminal nodes 5 and 6 show the opposite expression structure, indicating a marked difference in profiles at these nodes. This fits with their relative positions within the tree.

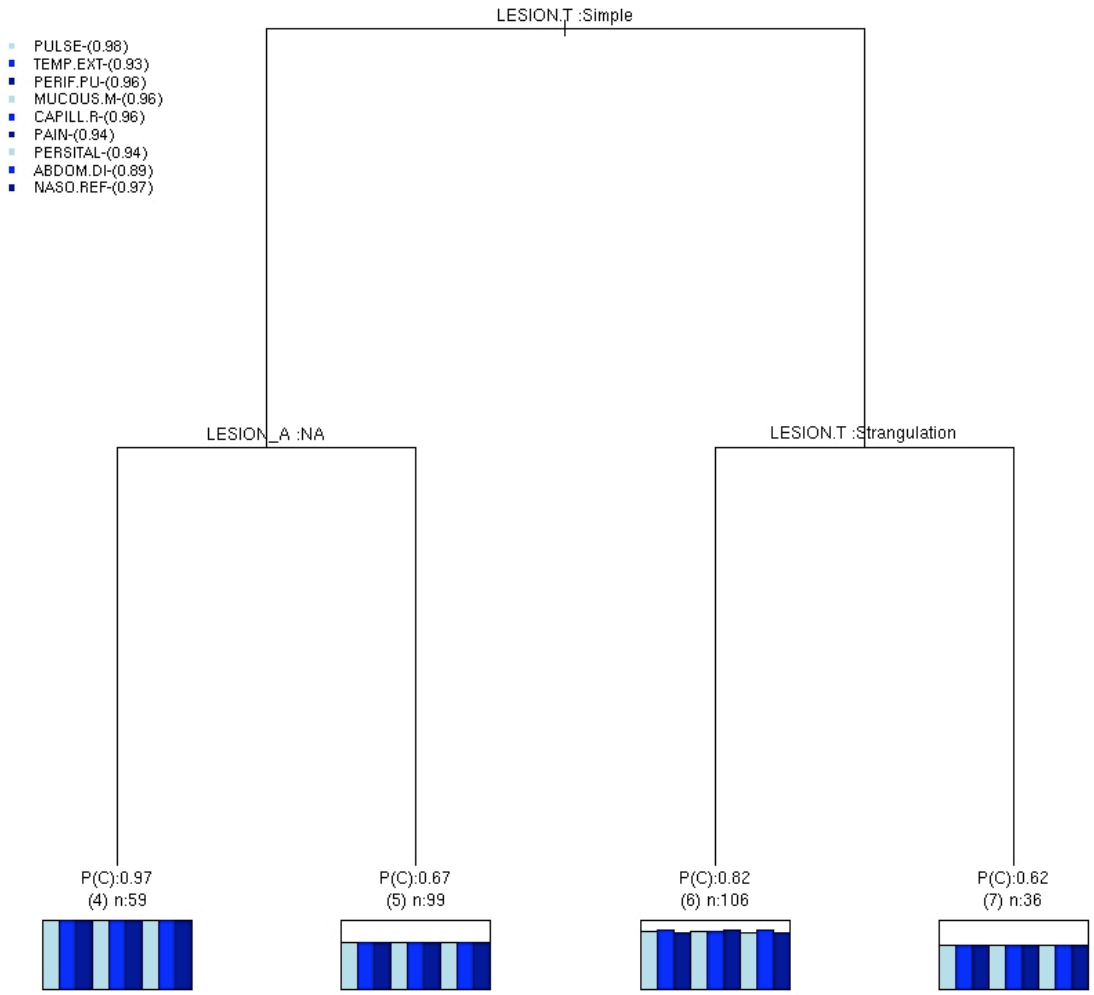
Table 28: Horse colic analysis plaid terminal node plaid filtering results.

Response Variable	MCT Node 4	5	6	7
PULSE	0	0	4.57	0
TEMP.EXT	0	6.04	0	0
PERIF.PU	0	0	-6.92	-11.40
MUCOUS.M	0	5.03	0	0
CAPILL.R	17.03	-4.27	0	0
PAIN	0	6.99	0	0
PERISTAL	0	0	-3.57	0
ABDOM.DI	-17.29	0	5.79	-11.81
NASO.REF	0	-13.78	0	23.09

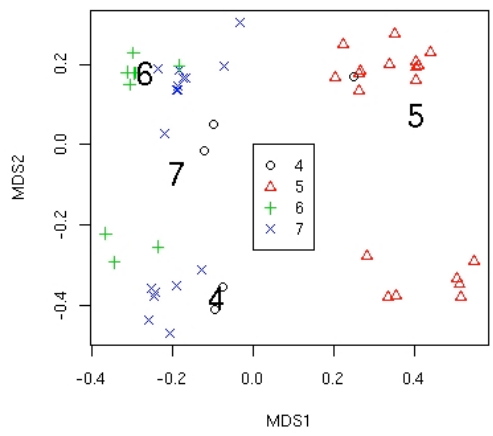
Terminal node filtering offers a means to test the homogeneity of each terminal node and investigate any variables that violate this assumption. However the MCT is built using information from all response variables, whether they are homogeneous with the MCT groups or not. It is possible that in a sufficiently complex response set that there will be sub-groupings of the variables that show different structure. We now propose an extension to the plaid combining method aimed at identifying these sub-groups before an MCT is build.

Figure 97: Horse colic analysis complete response set global MCT and terminal node location MDS plot.

(a) MCT



(b) Terminal node location MDS plot



7.3.3.3 Plaid response variable filtering algorithm

Plaid model RFP combination estimates a binary variable κ_m, which flags those RFPs whose configurations deviate from the mean configuration. An RFP with a κ_m of '1', has a different configuration from the mean, whereas a κ_m of '0' is considered to be adequately modelled by the background mean. Using a recursive algorithm described in Figure 98 it is possible to construct a search for similar configurations, by identifying those RFPs with κ_ms of '0'. The algorithm is stopped either when all κ_m s are either '1' or '0', or when the residual sums of squares between the RFPs and the combined configuration has converged.

If the residual sums of squares of the plaid model have converged, but there are still some κ_ms of '1', then plaid models considers these RFPs to be different but the effect of their difference is small. Therefore removing them does not improve the error in the modelled consensus structure. At this point, the RFPs are considered to be sufficiently homogeneous.

If the algorithm returns a subset where all κ_m s are found to be '1', it implies that all RFPs are sufficiently different from their background mean. Therefore no simple mean of the RFPs can be used to model the overall structure. If this occurs it is likely that PLAID combining will not yield the most accurate consensus matrix. In this case, a different combination method, designed to model heterogeneity between RFPs, such as the BB or GPA combination methods, should be employed to estimate the consensus.

Figure 98: Plaid variable filtering algorithm.

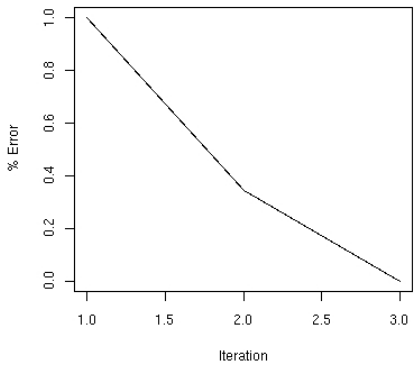
1. Place all RFPs in subset A.
2. **While** subset A has RFPs within it **do**:
    a. Calculate the complete plaid model parameters for all RFPs in subset A as described in Section 3.5.3.
    b. Compute the plaid model error, Q_i, by (3.38).
    c. Compute the percent error relative to the error in the plaid model involving all RFPs in the initial subset A, Q_0.
    d. **If** all $\kappa_m = 0$ **then** stop, a good subset of RFPs has been found.
    e. **If** all $\kappa_m = 1$ **then** stop, the RFPs cannot be modelled well by a stable mean representation.
    f. **If** the percent error has converged but all κ_m*'s* do not equal 0 **then** stop, a reasonable subset of RFPs has been found.
    g. Update subset A with all RFPs with a $\kappa_m = 0$.
    h. Update subset B with all RFPs with a $\kappa_m = 1$.
3. Rerun the analysis on subset B.

The result of the plaid variable filtering algorithm in Figure 98is 4 groups of response variables shown in Table 29. For all variables but PERIF.PU the RMSE error between the RFP and the consensus configuration reduced, sometimes by over a half. Furthermore the variable groups found appear to make physical sense, with group 1 and 2 relating primarily to the horse's blood circulation function and group 3 relating to any observed pain the horse may be experiencing. Finally group 4 just contains MUCOUS.M (a variable describing the colour of the horse's eyes) and is grouped separately as it does not obviously relate to either heart function or observed pain.

The first group is found with some κ_ms not being zero. This is because the plaid model error is shown to be sufficiently small at two iterations (Figure 99). If the filtering algorithm is followed through to the third iteration, the first group of RFPs only contains TEMP.EXT. The results in Figure 99 show that at iteration 2 the RFPs contribute to less than 34 % of initial plaid model error at iteration 1. Because the error decrease from the second to the third iteration is small, the RFPs with a κ_m of '1'

are shown to have a minimal effect on the homogeneity of the final consensus. Therefore they are considered sufficiently modelled by the consensus in the second iteration and the first group of RFPs is defined to be TEMP.EXT, PERIF.PU and CAPILL.R.

Figure 99: Horse colic analysis plaid variable reduction error convergence for the first group.



To investigate any improvements in resolution over these subgroups a global MCT is now built upon them. For the first group, the RE curves (Figure 101a) indicate a tree size of three or four splits is possible. This is one more split than the full response set MCT. In this MCT (Figure 101b) the first three splits are the same as the full MCT, and the additional split acts upon the full MCT's terminal node 5. The resulting terminal nodes 10 and 11, improve the probability of expression from 0.65 in the full MCT to approximately 0.77 and 0.72 respectively, in the reduced MCT. The MDS plot of the terminal node groups (Figure 101c) is more clearly resolved than in the full MCT with the noticeable difference in the group separation.

By the RE curve for the second group (Figure 102a), 3 splits are selected. The resulting MCT (Figure 102b) is identical to the full MCT in the splitting structure.

The only difference being in the MDS plot (Figure 102c), which appears to more clearly identify group 4.

The third group RE curve (Figure 103a) clearly indicates 2 splits, a smaller tree to the other groups. The splits made (Figure 103b) are the same as in the full MCT tree, however it does not make the partition to find nodes 7 and 8. The MDS plot (Figure 103c) shows a clear separation between nodes 3 and 5, however node 4 is not easily identified.

As the fourth group only has one variable, a consensus matrix does not need to be computed. The RE curves (Figure 104a) for this group indicate 4 splits as in the first group (Figure 104b). However by observation of the MDS plot (Figure 104c) the differences in the group structure between the two are apparent. Furthermore the fourth group more accurately defines groups 10 and 11. This is shown in the mean probability of expression within these terminal nodes increasing from 0.77 and 0.72 in the first variable set to 0.8 and 0.8 in the fourth set.

For each variable a strongly significant group profile is found (Figure 100). For categorical variables this was tested using a χ^2 test of independence between the group categories and the MCT terminal node labels, and for a continuous variable a one-way ANOVA was performed testing the mean difference between MCT terminal nodes. The significance of these profiles indicates the groups found by the MCTs are representative of structure within the responses. Correlation coefficients, r, are also

presented to assess the strength of the relationships, for continuous variables Pearson's r is computed and for categorical variables Cramer's Phi is computed.

From the response variable profiles (Figure 100) it can be seen that the group structure found is weak. Group 1 appears to be defined by the temperature at the horse's extremities being either normal or reduced, a reduced pulse and a capillary refill time of less than 3 seconds. Group 2 finds MCT terminal nodes 4, 5 and 6 relating to no nasogastric reflux. The significant difference seen over the terminal nodes is driven by an elevation in pulse between nodes 4 and 7. In fact all of these profiles significantly highlight terminal node 4 as showing different structure. Terminal node 4 in these groups more often identifies the groups labelled 'normal' within the response variables. As a result an overall interpretation of these results is that RFP groups 1 and 2 are be primarily focused on identifying the profiles of a normal horse.

Group 3 has the strongest observed correlations however no clear group structure exists over the variables. Reversing the problem to a classification problem discriminating the groups in the predictor variable LESION, it is seen that only group 3 variables are used in the tree (Figure 105). This indicates the response variables within this group are those that are highly predictive of a single response variable LESION and therefore are determining the dominant grouping structure within the dataset.

Table 29: Horse colic analysis plaid response variable filtering results.

1. Group one

Variables:	TEMP.EXT	PERIF.PU	CAPILL.R
κ_m:	0	1	1
β_m:	0	-5.56	5.56
RMSE with filtered response consensus:	8.15	8.90	4.85
RMSE with full response set consensus:	16.26	8.49	10.41

2. Second Group

Variables:	PULSE	NASO.REF
κ_m**:**	0	0
β_m**:**	0	0
RMSE with filtered response consensus:	6.26	6.26
RMSE with full response set consensus:	18.90	10.37

3. Third Group

Variables:	PAIN	PERSITAL	ABDOM.DI
κ_m:	0	0	0
β_m:	0	0	0
RMSE with filtered response consensus:	4.59	6.58	6.49
RMSE with full response set consensus:	8.33	9.52	10.31

4. **Fourth Group**: MUCOUS.M

Figure 100: Horse colic analysis response variable group profiles.

(a) Group 1

TEMP.EXT

	MCT Node				
	4	**6**	**7**	**10**	**11**
Absent	2	17	3	5	0
Reduced	10	46	19	24	10
Increased	11	6	1	10	2
Normal	25	18	5	24	6

$\chi^2 = 43.044$
r = 0.19
P-Value = 0.003

PERIF.PU

	MCT Node				
	4	**6**	**7**	**10**	**11**
Absent	0	4	2	2	0
Reduced	3	49	19	23	9
Increased	3	1	0	1	0
Normal	39	25	8	34	9

$\chi^2 = 53.7028$
r = 0.21
P-Value = 0.00009

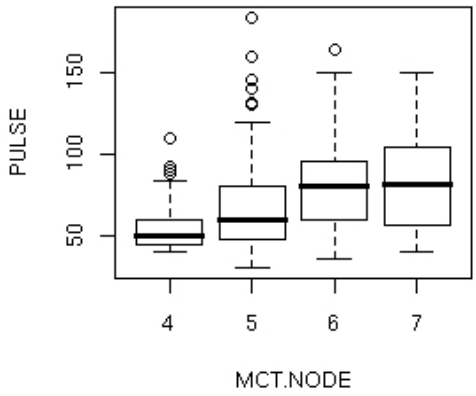
CAPILL.R

	MCT Node				
	4	**6**	**7**	**10**	**11**
>= 3 Seconds	0	41	15	14	5
< 3 Seconds	48	57	15	56	12

$\chi^2 = 30.0507$
r = 0.16
P-Value = 0.00009

(b) Group 2

PULSE



$F = 36.908$
r = 0.36
P-Value = $4.128 * 10^{-9}$

NASO.REF

	MCT Node			
	4	**5**	**6**	**7**
< 1 Litre	1	12	20	7
> 1 Litre	0	12	15	8
None	27	43	44	6

$\chi^2 = 25.6809$
r = 0.17
P-Value = 0.0005

(c) Group 3

PAIN

	MCT Node		
	3	**4**	**5**
Continuous severe pain	31	1	10
Intermittent severe pain	25	1	13
Intermittent mild pain	29	13	25
Depressed	31	10	18
Alert, no pain	2	23	13

$\chi^2 = 69.2889$
r = 0.34
P-Value = 0.00009

PERISTAL

	MCT Node		
	3	**4**	**5**
Abset	51	0	22
Hypomotile	56	23	49
Normal	4	6	6
Hypermotile	8	23	8

$\chi^2 = 65.4187$
r = 0.33
P-Value = 0.00009

ABDOM.DI

	MCT Node		
	3	**4**	**5**
Severe	23	0	15
Moderate	47	2	16
Slight	26	17	22
None	19	27	30

$\chi^2 = 50.2653$
r = 0.29
P-Value = 0.00009

(d) Group 4

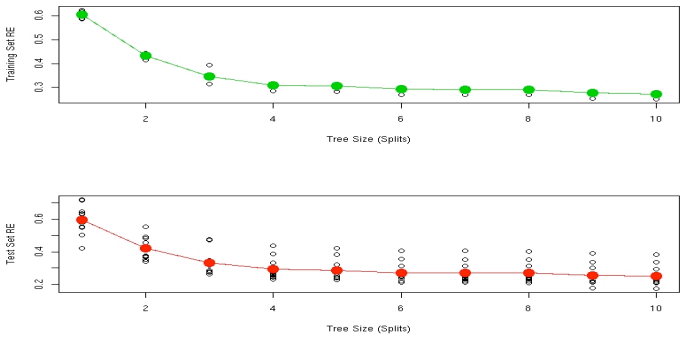
MUCOUS.M

	MCT Node				
	4	**6**	**7**	**10**	**11**
Dark Cyanotic	1	11	6	2	0
Bright Red/ Injected	2	6	10	5	2
Pale Cyanotic	1	28	4	8	0
Pale Pink	7	22	3	20	6
Bright Pink	10	10	2	5	3
Normal Pink	29	13	6	15	16

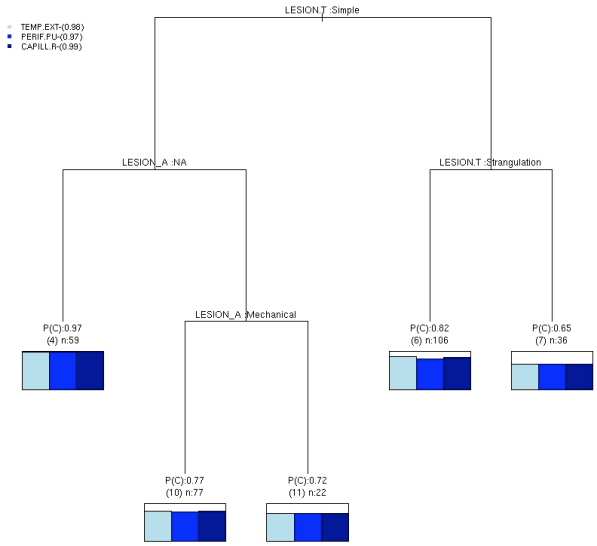
$\chi^2 = 95.29$
r = 0.28
P-Value = 0.00009

Figure 101: Horse colic analysis plaid filtered variable group one MCT results.

(a) 10 fold CV RE curve



(b) MCT



(c) Terminal node location MDS plot

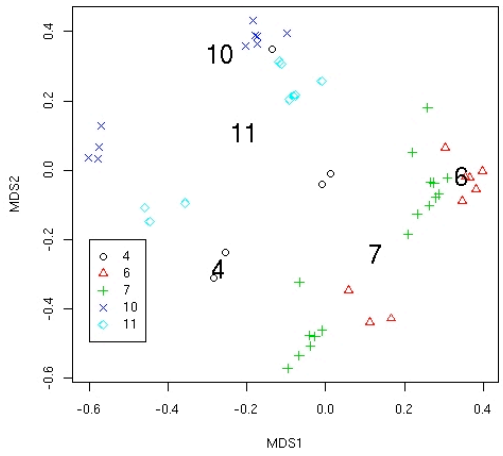
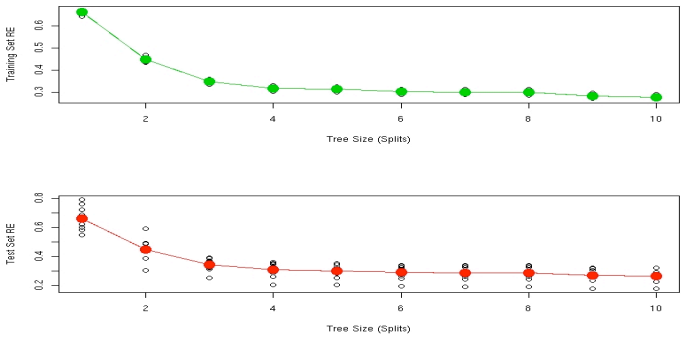
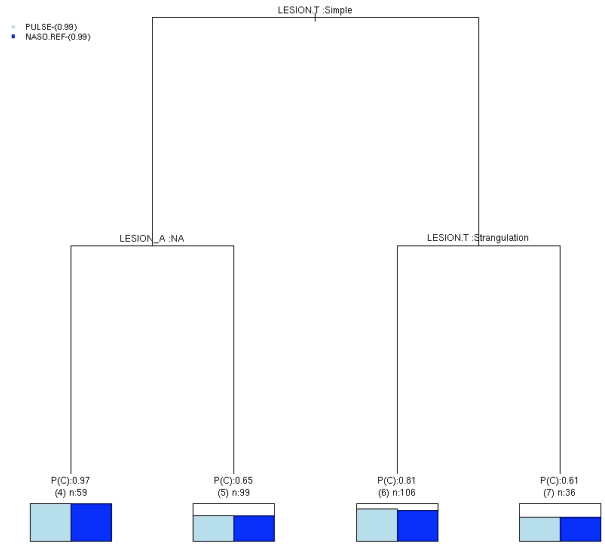


Figure 102: Horse colic analysis plaid filtered variable group two MCT results.

(a) 10-fold CV RE curve



(b) MCT



(c) Terminal node location MDS plot

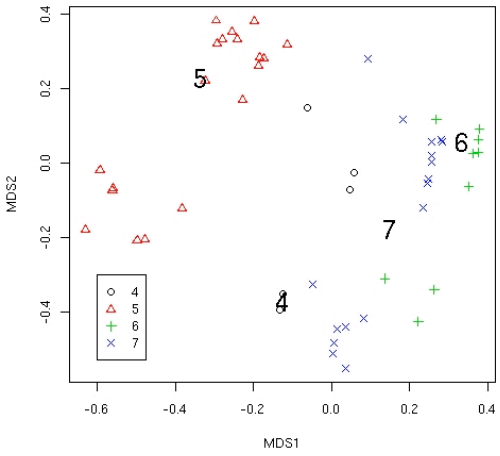
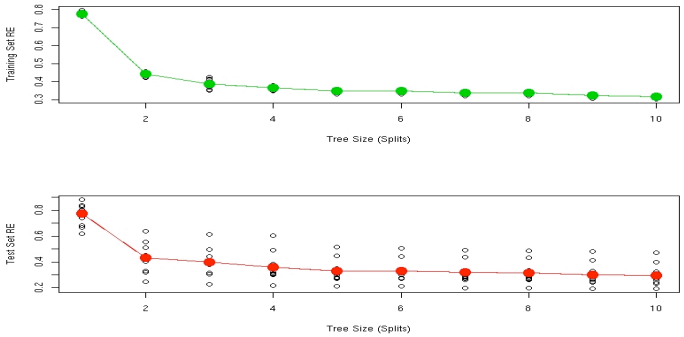
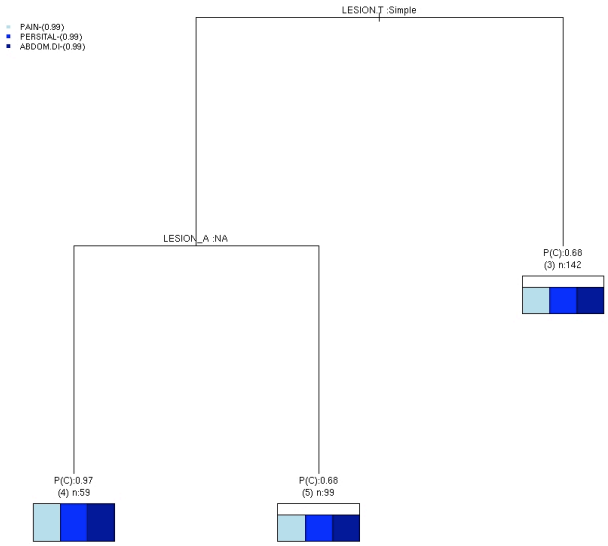


Figure 103: Horse colic analysis plaid filtered variable group three MCT results.

(a) 10-fold CV RE curves



(b) MCT



(c) Terminal node location MDS plot

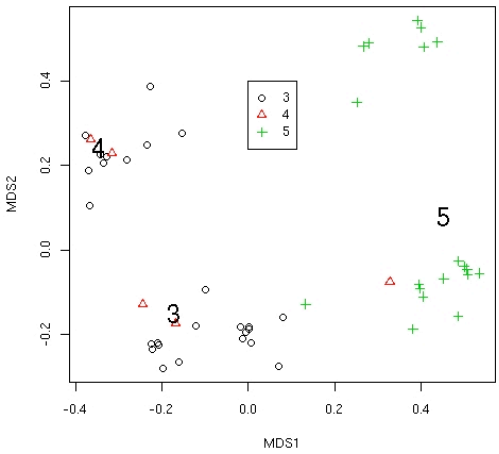
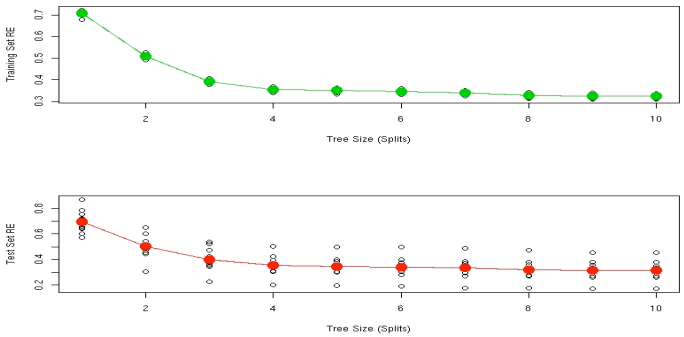
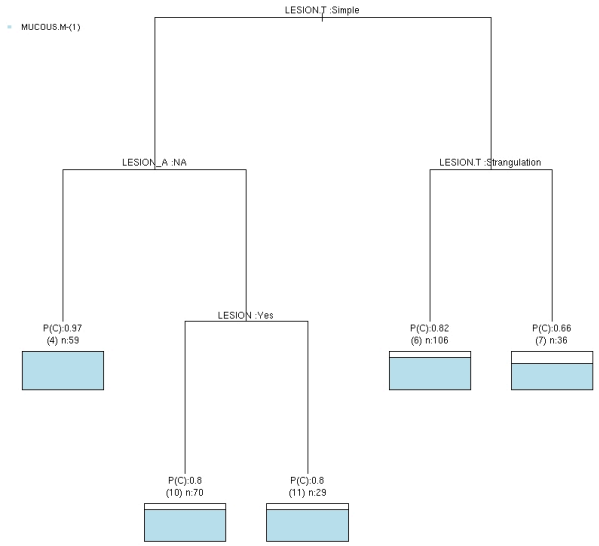


Figure 104: Horse colic analysis plaid filtered variable group four MCT results.

(a) 10-fold CV RE curve



(b) MCT



(c) Terminal node locations MDS plot

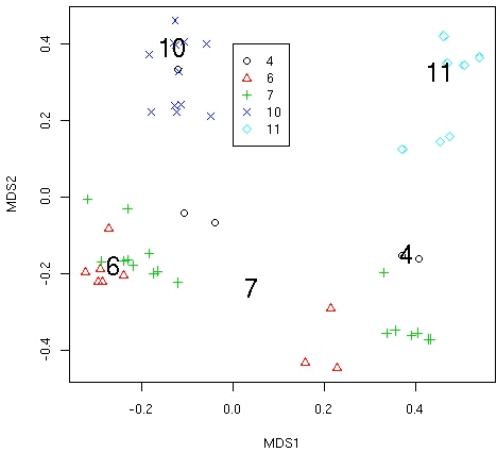


Figure 105: Horse colic analysis classification tree classifying the groups within predictor LESION by the entire response set. (Correct classification rate of 77.33 %).

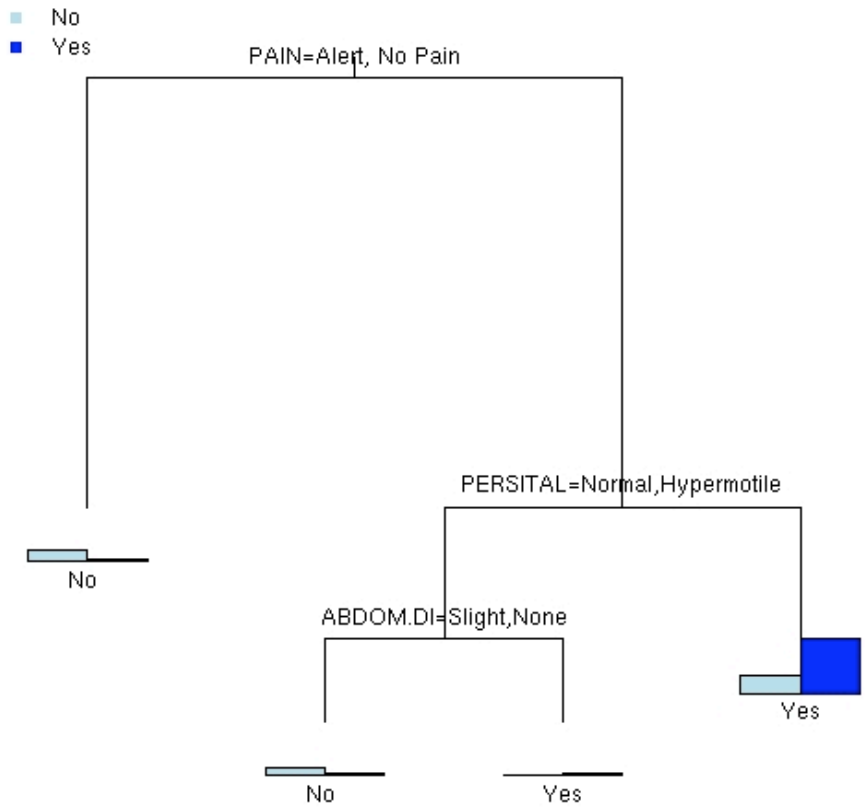
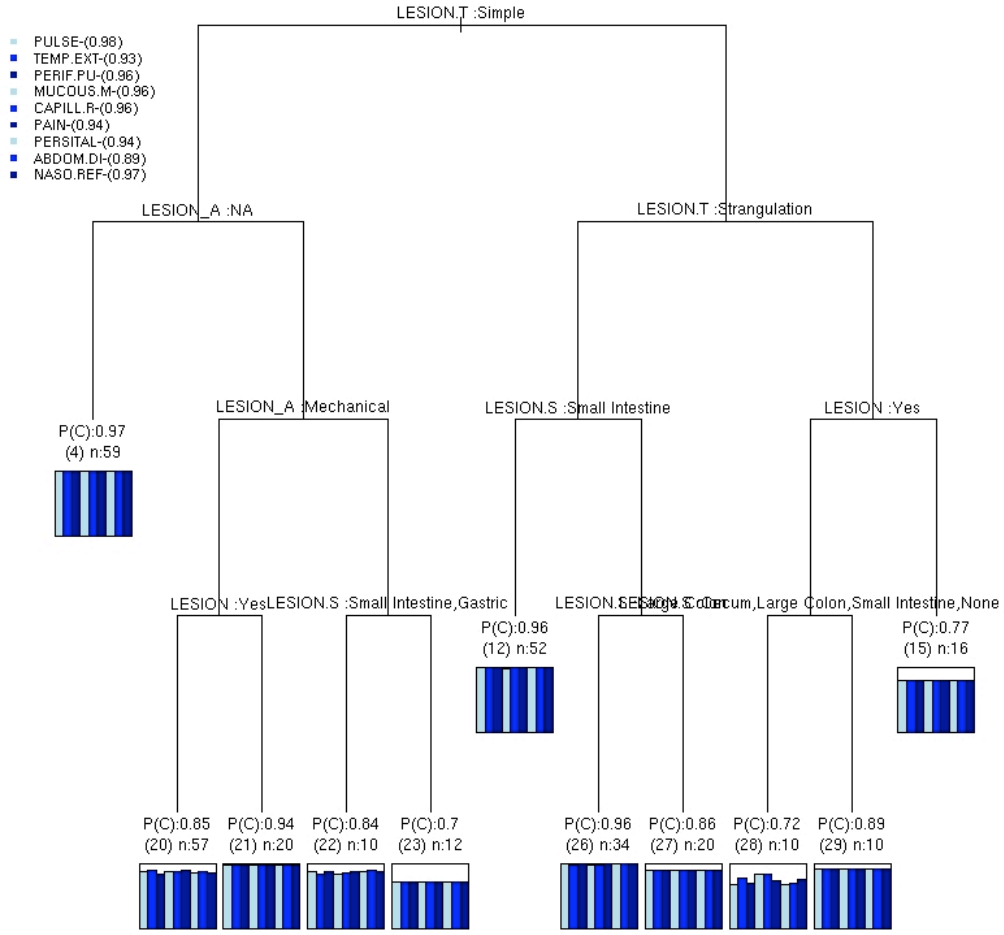
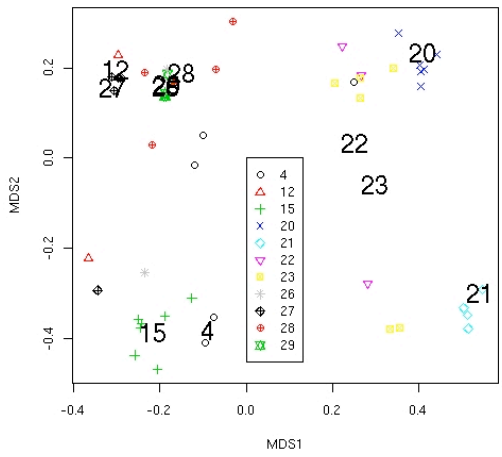


Figure 106: Horse colic analysis MCT grown to 10 splits.

(a) 10 split MCT



(b) 10 split MCT terminal node locations



7.4 Horse Colic Summary

The horse colic dataset is a good example of a profiling style analysis with MCTs. Firstly, using the a simple MRT on either the Gower distance matrix or binary substituted representation of the profiling set resulted in poor results. The improvement gained from moving to the random forest and treeboost proximity matrix is considerable. Groups that are common to both predictor and profiling sets now become obvious and easily found using the MCT splitting criteria. These methods give good indications of the structure to be found within the analysis however provide little detailed information on the composition of the groups.

Using MCTs it is possible to observe the grouping structure of each individual response variable and the relationship with the consensus matrix. It allows for the terminal nodes of the resulting MCT to be simplified using plaid filtering. By using MCTs with plaid terminal node filtering a two-way clustering is performed, where within a terminal node lie a subset of response variables and observations that define the common profile within the group.

A pre-processing step can be taken with the recursive filtering algorithm, allowing for an initial clustering of the responses based on the structure within their RFPs. This analysis highlights the complexities within profiling studies, as each response group displays a different subset of groups. What is interesting in this analysis is not the differences but the similarities between the subsets: in this case the splitting variables used in the tree and the tree size. It is clear that modelling a subset of variables produces a more accurate result, however this improved accuracy relates to the same

groups found in the overall analysis. It did not change the consensus structure completely, but reinforced the groups found in the overall consensus.

A major issue with the horse colic analysis was the high level of missing values. The results for the plaid filtering are dependant on the original global MCT model shown in Figure 97 and therefore if any bias in the missing values exists it will be obvious in this model. The percent of missing values in each terminal is shown in Table 30 show that terminal nodes 5 and 6 contain the 69 % of missing values and 4 and 7 only contain 31 %. However comparing this distribution to the relative size of each terminal node it is seen that the missing values are distributed with terminal node size. Therefore no obvious bias towards any particular group of missing values is observed.

Table 30: Horse colic analysis global MCT terminal missing value distribution.

MCT Node	**4**	**5**	**6**	**7**
% Missing values	0.19	0.34	0.35	0.12
Relative terminal node size	0.22	0.33	0.35	0.12

8. Discussion

The aim of this thesis is to extend tree based methods to handle a mixed type multivariate response. To do this a series of methods have been developed. Firstly, mixed type extensions to a multivariate tree are implemented by transforming the response using either the Gower distance, or binary substitution. These techniques offer a simple solution to a complex problem, but provide little in the way of understanding the result. Secondly, to improve on the performance of a single tree, multivariate tree based ensemble methods are also developed. Ensemble methods improve the predictions on the multivariate responses, and by binary substitution, are further extended to mixed type response sets. Multivariate tree-based ensembles are shown in this thesis to be powerful methods for profiling.

One key feature provided by tree-based ensemble methods is their proximity matrices. These proximity matrices are identical to consensus matrices that can be produced over a cluster ensemble. This changes the interpretation of a tree-based ensemble to that of a consensus clustering algorithm. A result of this interpretation is that the predictive performance of the ensemble becomes a key statistic in determining the quality of the final clustering solution. By using the ensemble predictive performance the problems in determining the accuracy and reproducibility of a cluster ensemble are reduced.

The major contribution of this thesis is the development of multivariate consensus trees (MCT) for mixed type clustering or profiling. MCTs combine ensemble proximities into one overall consensus matrix in an analogous step to the cluster ensemble search for the overall partition. This provides more information on the accuracy of the final solution with the ability to analyse the individual group structure

of each response variable in the analysis. MCTs partition the consensus matrix to find the optimal partition. This procedure uses decision rules to map the predictor variables back over the consensus to allow for an understanding of the origin of the final groups in the optimal partition. These rules also make MCTs a predictive clustering or profiling algorithm allowing them to easily group new observations without altering the original model. This predictive ability allows MCTs to cross-validate estimates on the number of groups and overall group accuracy.

Before opting for the more complex and computationally expensive solution as implemented in MCTs, using the simple tree and ensemble methods can be useful. The Gower distance metric and binary substitution transformation of the response set are common ways of finding groups in mixed type domains. In this thesis the results of these approaches are remarkably similar to each other as they both assume a Euclidean relationship between categorical and quantitative variables. This similarity is highlighted in the breast cancer dataset analysis. In this example both approaches grow the same tree and the MDS plots show very similar group structure (Figure 65, Figure 67). Binary substitution is the more flexible of the two approaches as it can also be used with ensemble tree methods. In the case of obvious structure these simple extensions will work.

A major problem for binary substituting of the response is that of dimensionality. Binary substitution inflates the number of variables within the response by the total number of levels within each categorical variable. In the case of the breast cancer analysis the 9 original categorical variables were transformed into 89 binary variables. Although it has been shown that multivariate tree based methods can handle a large

response set (Smyth, Coomans and Everingham 2006), the understanding of the final result is impaired by the dimensionality. Furthermore as the response set is treated as a whole, filtering out unimportant variation is not possible.

Multivariate extensions to tree based ensembles are shown to clearly improve group resolution within the MDS plots of the proximity matrices. On comparison between multivariate random forest and treeboost notable differences in the group structure of the responses are observed. The group structure within the random forest proximity matrices more closely matches that observed using the Gower distance and binary substitution. However treeboost appears to find a consistently different group structure as seen in the thyroid and breast cancer analyses. This difference does not manifest itself in performance, with the final grouping of the treeboost proximity matrix outperforming the final random forest proximity.

Despite the improved accuracy observed when determining the groups over the treeboost proximities, they are not appropriate inputs for the MCT consensus construction. There are two reasons for this:

1) Boosting models are sensitive to the shrinkage parameter. This prohibits automated running of the model as required for MCT construction. The action of the shrinkage parameter means that simply increasing the number of trees within the model will not achieve optimal performance (Hastie, et al. 2001). Random forests however can be easily tuned by increasing the number of trees within the forest to achieve optimal performance, and because of this are ideal candidates for MCT construction.

2) The trees in a boosting model are dependant upon each other. This means that each tree does not contribute equally to the construction of the proximity, a fact that is not reflected within the proximity matrix itself. This violates the assumption of a binomial distribution of the counts and could seriously affect the combination methods.

The analysis of the consensus matrix with the MDS plots must go hand in hand with a heat map of the reordered matrix. The structure of the groups with the MDS plots does not represent their structure within the dataset but how well that group has been predicted by the ensemble. The result of this is that groups that are poorly predicted will be large and noisy within the MDS plot. These groups will also have a relatively low probability of expression.

From the base tree and ensemble methods it is clear that trees are highly suited to mixed type clustering and profiling. The primary feature of tree-based methods is the ensemble proximity matrices. By partitioning these matrices it is possible to simultaneously view a logical decision path that predicts each group in the form of a tree and the relationships between these groups within the MDS plots, a feature that is not available with any other unsupervised technique. This allows for a detailed understanding on how the groups within the predictor set match the response set. However as the response set is treated as a whole, they do not allow for clear understanding of how well each response variable expresses each group. To do this the more individualistic analysis of MCTs is required.

MCTs are designed for simultaneous analysis of relationships in both the response variables and between the response and predictor variables. This is done by

individually analysing the group structure within each response variable. By combining these structures MCTs not only retain all the functionality of the random forest proximity matrices but also can improve on the group resolution. In fact this thesis shows the performance of MCTs for unsupervised classification can be comparable to the performance of a classification tree. Also by analysis of the individual RFP structures it is possible to filter noise and unrelated variables from the response set by using both performance diagnostics and plaid combining.

By extending PLAID combining, MCTs offer an algorithm to filter the response variable set. Plaid filtering is implemented in two ways, firstly to cluster the response variables before construction of an MCT, and secondly to test the assumption of homogeneity within the terminal nodes of an existing tree. Plaid filtering extends MCTs to be a two-way technique, where a group is defined both on a subset of observations and variables. In the horse colic analysis plaid filtering is used to cluster the variables within the response variable set. Over the four response variable subgroups found, different group structure within them is observed. Furthermore, the consensus produced from each subgroup is a more accurate consensus in terms of RMSE between RFPs of the subgroup and the overall model consensus.

The horse colic results give a clear indication of the power of plaid filtering. Firstly the algorithm removes the dominant variation corresponding to the normal symptoms of a horse, and then places together the variables related to the lesion groups. In addition the profiles found for each subgroup are different and also strongly significant. However the relationships in terms of correlation observed over the

terminal nodes are weak ranging between 0.16 to 0.36. Therefore plaid filtering is shown to be effective even when the observed grouping structure is weak.

However, as plaid filtering is finding groups over RFPs generated from a random procedure, care should be taken to ensure that all structure in these matrices has been fully resolved. This can be done easily by increasing the number of trees within each forest and observing the structure change. If the RFPs themselves are unstable, then plaid filtering will also be unstable.

Much of the effort in this thesis is spent of testing the effect of various parameter specifications in the three stages of MCT construction (Table 2). The estimation of the overall consensus matrix from the individual RFPs is the first major complexity within the MCT algorithm. Three combination methods are proposed in this thesis, GPA, BB and PLAID. Both GPA and PLAID define the overall consensus by minimising the square error loss between each individual RFP and the consensus. BB does not minimise a loss function but rather provides a robust estimation of each count within the proximities by estimating their overall probability distribution. As a result it is expected that different combination methods will provide a different consensus solution.

By analysing the RMSE errors between the RFPs and the consensus matrix it is possible to assess the quality of the combination. This provides a response variable importance statistic for the overall MCT. Strong similarity is found in the resulting consensus matrices from each combination method. From the results it appears that GPA and BB are finding very similar structure as they produce the same global and

local MCT in the breast cancer analysis (Figure 79, Figure 81) and show the same performance convergence in the sensitivity analysis (Table 16, Table 17). PLAID combining however produces a different global MCT for the breast cancer dataset (Figure 79), shows a different performance convergence in sensitivity analysis (Figure 79, Figure 81) and shows a much increased RMSE for the 10 uneven but clear group simulation tests (Figure 39). However whether these differences translate into reduced clustering accuracy is not clear. In the sensitivity analysis using PLAID combining shows a less accurate consensus that resulted in a reduced performance of the overall tree. However in the breast cancer dataset, using PLAID combining results in the most accurate tree.

The inconsistent performance of PLAID combining is most likely due to the plaid model's search for common structure over the RFPs. In this thesis the plaid model is only run to a single layer. This may result in smaller groups being modelled in later layers, as the common structure in the first layer is likely to favour the larger groups. This is what is observed in the sensitivity analysis. With plaid models the MCTs grown using the PLAID consensus do not finding the smallest group (group 3) and in the ten group simulation experiment with large and small group sizes the RMSE for plaid combining is obviously the greatest. These results show that PLAID combining may not resolve smaller group structure over the RFPs.

An obvious solution to this problem is running plaid combining to more than one layer. However, as different group configurations will exist within each additional layer any interpretation of what variables contribute to the groups in the final consensus will be confused. This removes one of the most important features of the

plaid combining algorithm. Another approach is to change the background layer from an average consensus matrix to one produced by BB or GPA. This approach biases plaid models to the structure found by BB and GPA. A flow on affect is to change the interpretation of the plaid parameters. Instead of modelling the deviations from the mean consensus they are modelling the deviations from a modelled consensus. As this modelled consensus has no simple expression, the interpretation of the plaid parameters as estimating deviations from homogeneity does not hold.

This thesis also assessed the effect noise variables within the RFPs will have on the final consensus solution. The results showed that the consensus generation procedure of MCTs was found to be remarkable resistant to added noise within the response set. In these experiments it is shown that the consensus configuration has the same group structure as the original despite the addition of pure randomness within the consensus generation procedure. Furthermore the RE curves accurately estimate the number of clusters, and the accuracy of the resulting partitions is found to be comparable or exceed that of K-means.

Once an overall consensus has been estimated the task is now to partition the matrix to find the groups. To do this five splitting criteria are developed. These criteria search over all decisions within each variable in the predictor for blocks of observations with high similarity within the consensus matrix. In an ideal case the decisions found will reorder the consensus matrix into a block diagonal structure where the similarities on the block diagonal are high and the similarities within the off diagonal blocks are low. Of the five splitting criteria developed, one observes the

group variance structure (SSR), two utilise the count structure of the cells within the matrix (MR and OR) and the last two are combinations of MR and OR with SSR.

The quality of each splitting criteria is best assessed by observation of their respective RE curves. Over the simulation tests these curves were produced for each criteria for each experiment. Overall it appears that MR and MR-SSR produced curves that are less accurate than the other splitting criteria (Figure 35, Figure 41, Figure 45). This is shown by consistent high variability within the cross-validated performances. The other splitting methods performed indistinguishably as the RE curves are closely matched and the misclassification performances similar. However when moved from the sterile domain of the simulation experiments to an actual dataset a clear interaction between the performance of the splitting criteria, combination method and random forest parameters emerged.

Considerable effort has been made in this thesis to quantify the interaction between the splitting criteria, combination method and random forest parameters for both global and local MCTs. Global MCTs have a clear interaction with random forest tree terminal node size. The structure of this is that if the terminal node size is set too small by increasing the tree size optimal performance can be reached (Table 16). This interaction is seen to be mostly independent of combination method. However local MCTs seemed only to be sensitive to terminal node size (Table 17). Here the terminal node size must be specified as close as possible to the smallest group size in the data.

These results are not surprising given how the two types of MCTs are grown. As local MCTs recompute the consensus matrix it is difficult to optimise the performance of the base random forests. Furthermore the choice of MCT splitting criteria is a sensitive parameter as the accuracy of the intermediate consensus matrices is dependant on the previous decisions within the tree. Global MCTs do not suffer as severely from this problem because the response is constant and therefore it is easier to optimise the important parameters. These differences are highlighted in the sensitivity analysis of the Vietnam data.

Local MCTs, when optimised, show a more improved resolution of the groups. This improvement highlights the power of a localised clustering solution. The resolution is improved as once obvious groups are removed from the analysis, more attention can be paid to separating the groups that are closer together. This is strongly highlighted in the breast cancer dataset analysis. To get the same performance the local MCTs require 3 splits whereas global MCTs require 5. This increased split accuracy is highlighted in the identification of the benign group. Global MCTs identify the majority of the benign group at terminal node 15, and much of the work in the early splits of the tree is dedicated to shaving off smaller malignant sub-groups (Figure 79). However local MCTs find the majority of the benign group first, in terminal node 3, and then use the other two splits to find the less obvious malignant sub-groups (Figure 81). This implies that local MCTs will find the most obvious groups first, whereas global MCTs are likely to favour smaller groups.

MCTs however are limited by their tree structure when finding groups. In the thyroid, Vietnam and breast cancer analyses it was found the performance of MCTs

approaches the performance of a classification tree. However in the thyroid analysis it was apparent that K-Means and PAM on the consensus matrix found by MCTs identified the known groups more accurately. For example, the local MCT for the thyroid dataset misclassifies 15 observations, whereas on the ACM, K-means misclassifies 12 and PAM 10 observations. The reason for this is that tree-based clustering methods are bound by groups that are separable by a single decision on a single predictor variable, where as K-Means and PAM are not. A possible solution to this is to define a multivariate or linear combination of splitting functions (Breiman, et al. 1984, Brodley and Utgoff 1995).

MCTs when run correctly are a powerful technique for clustering or profiling. However there are a lot of parameters that can serious affect the accuracy of the final solution. For a reasonable dataset as computation time for MCTs is considerable a course of action to determine a reasonable set of parameters for a MCTs analysis is now described:

1. The first step is to produce a single multivariate tree upon the dataset. If mixed types exist then use binary substitution or the Gower distance approaches. From this analysis it is hoped that the following information is gained:
    a. To determine appropriate tree and terminal node sizes for the ensemble methods. These can be determined through observation of the RE graphs.
    b. To assess the predictive performance of that tree. If there is no stable tree observed from the RE graphs then it is unlikely that this will change with any further analysis.
2. Once you have appropriate estimates for tree and terminal node size the next step is to build a simple multivariate random forest or treeboost model using these

parameters. From these models the most important piece of information is the error convergence plots. It is strongly recommended that a reasonable independent test set be used to assess the performance of the ensemble. If the predictions upon this test set do not converge then no stable trees can be built, and as a result this analysis will not find stable groups. If the performance does stabilise it is recommended that more trees well past the point of convergence be added to the model to ensure that this stability remains.

3. Once the ensemble methods are stable then the MCT approaches can be considered. Firstly observe the structure within the plots of the proximity matrix to get an idea of the quality of the group structure. Then produce the RE curves to partition this proximity matrix for each splitting criteria. If a stable group structure has been found the elbow should appear at the number of groups observed in the proximity matrix plots. If this is the case then MCT methods are likely to find a representative set of clusters.
4. The decision to go to the local or global MCT methods using a combined consensus matrix should be determined on the number of response variables available. If there are many responses then filtering out some before combining is recommended. In the horse colic analysis this was done on the basis of the performance of the random forest for each response variable. From here it is advisable to perform all combination techniques but only growing a global MCT on each. Once the global MCT parameters have been optimised, then a local MCT approach using similar parameters may be attempted. It should be noted that local MCTs are remarkably more sensitive than global MCTs to the choice of parameters and may take some time to optimise.

5. After the MCTs are grown and the results make sense, plaid filtering can be performed. However it is advisable that a stable model be found before performing this step as reproducible proximity matrices for each varaible are required.

It is hoped that this guide will produce stable MCT solutions, however the final result will be dataset dependent.

9. Conclusions

This thesis has developed models for mixed type clustering or profiling. The core idea within this thesis is that groups found to describe a dataset must be predictive of each variable within it. The methods developed in this thesis use tree-based ensemble techniques to predict the data, and cluster ensemble ideas to identify the overall grouping structure. This combination of ideas culminated in the development of a new algorithm called Multivariate Consensus Trees (MCT).

Multivariate Consensus Trees, in this thesis have been shown to find more accurate grouping structure than either hierarchical agglomeration, K-Means or PAM. Furthermore they enable an analysis of the found groups in terms of: "*which predictor variables determine the groups?*"; "*which response variables express these groups?*" and the probability that these groups are representative of the data. MCTs also allow for pre and post-processing steps, using plaid models, to filter out response variables that do not express the groups found by the MCT. These features of MCTs make them a unique tool for finding and understanding the grouping structure over a mixed type dataset.

The focus of MCTs is in finding groups on a multivariate mixed type response. However this thesis has also suggested methods for mixed type prediction using multivariate extensions to tree-based ensembles using binary substitution of categorical variables within the response dataset. Multivariate random forests and treeboost are new methods of predictive profiling analysis that can highlight grouping structure, but are more focused on creating an accurate predictive model. Tree based models are resistant to overfitting problems and can handle large datasets. These features are highly desirable for any multivariate model.

Overall this thesis has exploited the flexibility of trees in handling mixed data types and extended them to a predictive multivariate ensemble. Moving from prediction to clustering this thesis views a tree-based ensemble as a consensus clustering algorithm. The result is multivariate consensus trees, a tree based clustering and profiling tool for mixed data types.

10. References

Ana L. N. Fred, and Anil K. Jain. (2005), "Combining multiple clusterings using evidence accumulation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27, 835-850.

Ben-Hur, A., Elisseeff, A., and Guyon, I. (2002), "A stability based method for discovering structure in clustered data," *Pacific Symposium on Biocomputing*, 6-17.

Breiman, Friedman, Olshen, and Stone (1984), *Classification and Regression Trees*, Chapman and Hall.

Breiman, L. (1996a), "Bagging Predictors," *Machine Learning*, 26, 123-140.

Breiman, L. (2001), "Random Forests," *Machine Learning*, 45, 5-32.

Brodley, C. E., and Utgoff, P. E. (1995), "Multivariate Decision Trees," *Machine Learning*, 19, 45-77.

Burnham, K. P., and Anderson, D. R. (2002), *Model Selection and Multimodel Inference. A Practical Information-Theoretic Approach* (Second ed.), Springer-Verlag.

Buuren, S. V., and Heiser, W. J. (1989), "Clustering N objects into K groups under optimal scaling of variables," *Psychometrika*, 54, 699-706.

Carroll, J. D., and Chang, J. J. (1970), "Analysis of individual differences in multidimensional scaling via an N-way generalization of "Eckart-Young" decomposition," *Psychometrika*, 35, 283-320.

Coomans, D., Broeckaert, M., Jonckheer, M., and Massart, D. L. (1983), "Comparison of Multivariate Disciminant Techniques for Clinical Data - Application to the Thyroid Functional State," *Methods of Information in Medicine*, 22, 93-101.

De Jong, S. (1993), "SIMPLS: an alternative approach to partial least squares regression," *Chemometrics and Intelligent Laboratory Systems*, 18, 251-263.

De'ath, G. (2002), "Multivariate Regression Trees: A new technique for modelling species-environment relationships," *Ecology*, 83, 1105-1117.

De'ath, G., and Fabricius, K. E. (2000), "Classification and Regression Trees: A powerful yet simple technique for ecological data analysis," *Ecology*, 81, 3178-3192.

Dietterich, T. G. (2000a), "Ensemble methods in machine learning," *Lecture Notes In Computer Science*, 1857, 1-15.

Dietterich, T. G. (2000b), "An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, Boosting and Randomization," *Machine Learning*, 40, 139-157.

Dudoit, S., and Fridlyand, J. (2002), "A prediction-based resampling method to estimate the number of clusters in a dataset," *Genome Biology*, 3, 0036.0031-0036.0021.

Dudoit, S., and Fridyland, J. (2003), "Bagging to improve the accuracy of a clustering procedure," *Bioinformatics*, 19, 1090-1099.

E. Dusseldorp, and J. J. Meulman. (2001), "Prediction in medicine by integrating regression trees into regression analysis with optimal scaling," *Methods of Information in Medicine*, 40, 403-409.

Efron, B. (1979), "Computers and the theory of statistics: thinking the unthinkable," *SIAM Review*, 21, 460-480.

Esposito, F., Malerba, D., and Semeraro, G. (1997), "A Comparitive analysis of methods for pruning decison trees," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19, 467-491.

Esposito, F., Malerba, D., Semeraro, G., and Tamma, V. (1999), "The effects of pruning methods on the predictive accuarcy of induced decision trees," *Applied Stochastic Models in Business and Industry*, 15, 277-299.

Everitt, B. (1993), *Cluster Analysis*, London: Arnold Publishers.

Fern, X. Z., and Brodley, C. E. (2004), "Solving Cluster Ensemble Problems by Bipartite Graph Partitioning," *ACM International Conference Proceedings Series*, 69.

Fisher, R. A. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, 7, 179-188.

Fraley, C., and Raftery, A. (2002), "Model Based Clustering, Descriminant Analysis and Density Estimation," *Journal of the American Statistical Association*, 97, 611-631.

Freund, Y., and Shapire, R. (1997), "A decision-theoretic generalization of online learning and an application to boosting," *Journal of Computer Systems and System Sciences*, 55, 119-139.

Friedman, J. (1999), "Stochastic Gradient Boosting," *Technical Report Stanford University*.

Friedman, J. (2001), "Greedy Function Approximation: the gradient boosting machine," *Annals of Statistics*, 29, 1189-1232.

Friedman, J., Hastie, T. J., and Tibshirani, R. J. (2000), "Additive logistic regression: a statistical view of boosting (with discussion)," *Annals of Statistics*, 28, 337-374.

Friedman, J., and Popescu, B. (2003), "Importance Sampled Learning Ensembles," *Technical Report Stanford University*.

Friedman, J. H., and Popescu, B. E. (2005), "Predictive learning via rule ensembles," *Technical Report Stanford University*.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (1997), *Bayesian Data Analysis*, Chapman and Hall.

Gifi, A. (1990), *Nonlinear Multivariate Analysis*, John Wiley & Sons.

Giudici, P. (2003), *Applied Data Mining*, John Wiley & Sons.

Gower, J. C. (1971a), "A general coefficient of similarity and some of its properties," *Biometrics*, 27, 857-871.

Gower, J. C. (1975), "Generalized Procrustes Analysis," *Psychometrika*, 40, 33-51.

Gower, J. C., and Hand, D. J. (1996), *Biplots*, Chapman and Hall.

Hancock, T., "R package: 'mct'," (2006), timothy.hancock@jcu.edu.au

Hancock, T., Put, R., Coomans, D., Heyden, Y. V., and Everingham, Y. (2005), "A performance comparison of modern statistical techniques for molecular descriptor selection and retention prediction in chromatographic QSRR studies," *Journal of Chemometrics and Intelligent Laboratory Systems*, 76, 185-196.

Hartigan, J. A. (1975), *Clustering Algorithms*, New York: Wiley.

Hastie, T., Tibshirani, R., and Friedman, J. (2001), *Elements of Statistical Learning*, Springer.

Hastie, T., Tibshirani, R., Narasimhan, B., and Chu, G., "R contributed package: 'impute'," (2005), www.r-project.org

Hastie, T. J., Buja, A., and Tibshirani, R. J. (1995), "Penalized discriminant analysis," *Annals of Statistics*, 23, 73-102.

Hoerl, A. E., and Kennard, R. W. (1970), "Ridge Regression: Biased estimation for nonorthogonal problems," *Technometrics*, 13, 55-67.

Hong, N. T. (1997), "Trace Element Analysis With Application To Environmental Pollutants Studies In Vietnam," *Doctor of Philosophy Dissertation, Chalmers University of Technology and Goteborg University. S-412 96 Goteborg Sweden*.

Hornik, K., "R contributed package: 'clue'," (2006), www.r-project.org

Hubert, L., and Arabie, P. (1985), "Comparing Partitions," *Journal of Classification*, 2, 193-218.

Karypis, G., and Kumar, V. (1998), "A fast and high quality multilevel scheme for partitioning irregular graphs," *SIAM Journal of Scientific Computing*, 20, 359-392.

Kass, G. V. (1980), "An Exploratory Technique for Investigating Large Quantities of Categorical Data," *Applied Statistics*, 29, 119-127.

Kaufman, L., and Rousseeuw, P. J. (1987), *Clustering by means of medoids, in Statistical Data Analysis based on the L1 Norm*, ed. Y. Dodge, Amserdam: Elsevier.

Kaufman, L., and Rousseeuw, P. J. (1990), *Finding groups in data*, John Wiley & Sons.

Kavsek, B., Lavrac, N., and Ferligoj, A. (2001), "Consensus Decision Trees: Using hierarchical clustering for data relabelling and reduction," *Proceedings of the 12th European Conference on Machine Learning*, 2167, 251-262.

Larsen, D. R., and Speckman, P. L. (2004), "Multivariate regression trees for analysis of abundance data," *Biometrics*, 60, 534-549.

Lazzeroni, L., and Owen, A. (2002), "Plaid Models for Gene Expression Data," *Statistical Sinica*, 12, 61-86.

Lebart, L., Morneau, A., and Warwick, K. M. (1984), *Multivariate Descriptive Statistical Analysis: Correspondence analysis and related techniques for large matrices*, John Wiley & Sons.

Leisch, F., and Dimitriadou, E., "R contributed package: 'mlbench'," (2005), www.r-project.org

Lent, B., Swami, A., and Widom, J. (1997), "Clustering Association Rules," *Proceedings of International Conference on Data Enginneering*, 220-231.

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., and Hornik, K., "R contributed package: 'cluster' (v. 1.10.4)," (2006), www.r-project.org

Mangasrian, O. L., and Wolberg, W. H. (1990), "Cancer diagnosis via linear programming," *SIAM News*, 23.

Mclachlan, G. J., Bean, R. W., and Peel, D. (2002), "A mixture model-based approach to the clustering of microarray expression data," *Bioinformatics*, 18, 413-422.

Mcleish, M., and Cecile, M. (1989), "Horse Colic Dataset," *http://lib.stat.cmu.edu*.

Milligan, G. W., and Cooper, M. C. (1985), "An examination of procedures for determining the number of clusters in a dataset," *Psycometrika*, 50, 159-179.

Mitchell, M. (1998), *An Introduction To Genetic Algorithms* (1998 ed.), The MIT Press.

Monti, S., Tamayo, P., Mesirov, J., and Golub, T. (2003), "Consensus Clustering: A Resampling-Based Method for Class Discovery and Visualisation of Gene Expression Microarray Data," *Machine Learning*, 52, 91-118.

Pantazi, S., Kagolovsky, Y., and Moehr, J. R. (2002), "Cluster Analysis of Wisconsin Breast Cancer Dataset Using Self-Organising Maps," *MIE2002, Budapest, Hungry*.

Questier, F., Put, R., Coomans, D., Walczak, B., and Vander Heyden, Y. (2004), "The use of CART and multivariate regression trees for supervised and unsupervised feature selection," *Chemometrics and Intelligent Laboratory Systems*, 76, 45-54.

Quinlan, J. R. (1986), "Induction of decision trees," *Machine Learning*, 1, 81-106.

Quinlan, J. R. (1987), "Simplifying decision trees," *International Journal of Man-Machine Studies*, 27, 221-234.

Quinlan, R. (1993), *C4.5: Programs for machine learning*, San Mateo: Morgan Kaufmann.

R Development Team, "R: A language and environment for statistical computing," (2005), http://www.r-project.org/

Raftery, A. E., Madigan, D., and Hoeting, J. (1997), "Bayesian Model Averaging for Linear Regression Models," *Journal of the American Statistical Association*, 92, 179-191.

Rencher, A. C. (2002), *Method of Multivariate Analysis*, John Wiley & Sons.

Sain, S. R., and Carmack, P. S. (2002), "Boosting multi-objective regression trees," *Computing Science and Statistics*, 34, 232-241.

Schapire, R. E. (1990), "The strength of weak learnability," *Machine Learning*, 5, 197-227.

Schapire, R. E. (2002), "The boosting approach to machine learning: An overview," *In MSRI Workshop on Nonlinear Estimation and Classification, Berkeley, CA, Mar. 2001. http://stat.bell-labs.com/who/cocteau/ nec/ and http://www.research.att.com/~schapire/boost.html*.

Schapire, R. E., Freund, Y., Bartlett, P., and Lee, W. S. (1998), "Boosting The Margin: A new explanation for the effectiveness of voting methods," *Annals of Statistics*, 26, 1651-1686.

Schork, M. A., and Remington, R. D. (2000), *Statistics with Applications To The Biological and Health Sciences* (Third ed.), Prentice Hall.

Seber, G. A. F. (1984), *Multivariate Observations*, John Wiley & Sons.

Segal, M. R. (1992), "Tree-structured methods for longitudinal data," *Journal of the American Statistical Association*, 87, 407-418.

Seong Keon Lee, Hyun-Cheol Kang, Sang-Tae Han, and Kwang-Hwan Kim. (2005), "Using Generalized Estimating Equations to Learn Decision Trees with Multivariate Responses," *Data mining and knowledge discovery*, 11, 273-293.

Shi, T., and Horvath, S. (2003), "Using random forests similarities in unsupervised learning: Applications to microarray data," *Atlantic Symposium on Computation Biology and Genome Informaticcs (CBGI'03)*.

Shi, T., and Horvath, S. (2006), "Unsupervised Learning with Random Forest Predictors," *Journal of Computation and Graphical Statistics*, 15, 118-138.

Smyth, C., Coomans, D., and Everingham, Y. (2006), "Clustering Noisy Data In A Reduced Dimensional Space via Multivariate Regression Trees," *Pattern Recognition*, 39, 424-431.

Smyth, C., Coomans, D., Everingham, Y., and Hancock, T. (2005), "Auto-associative Multivarite Regression Trees for Cluster Analysis," *Chemometrics and Intelligent Laboratory Systems*, 80, 120-129.

Srikant, R., and Agrawal, R. (1997), "Mining Generalized Association Rules," *Future Generation Computer Systems*, 13, 161-180.

Stephen Swift, Allan Tucker, Veronica Viniotti, Nigel Martin, Christine Orengo, Xiaohui Liu, and Paul Kellam. (2004), "Consensus clustering and functional interpretation of gene-expression data," *Genome Biology*, 5(11):R94.

Strehl, A. (2002), "Relationship-based Clustering and Cluster Ensembles for High-dimensional Data Mining," *Doctor of Philosophy Dissertation; The University of Texas*.

Strehl, A., and Ghosh, J. (2002), "Cluster Ensembles - A Knowledge Reuse Framework for Combining Multiple Partitions," *Journal of Machine Learning Research*, 3, 583-617.

Therneau, T. M., Atkinson, B., and Ripley, B., "R contributed pacakge: 'rpart'," (2005), www.r-project.org

Therneau, T. M., Atkinson, B., Ripley, B., and De'ath, G., "R contributed package: 'mvpart'," (2004), www.r-project.org

Tibshirani, R. J., Walther, G., Botstein, D., and Brown, P. (2005), "Cluster Validation by Prediction Strength," *Journal of Computation and Graphical Statistics*, 14, 511-238(518).

Tibshirani, R. J., Walther, G., and Hastie, T. J. (2001), "Estimating the number of clusters in a dataset via the gap statistic," *Journal of the Royal Statistical Society B*, 63, 411-424.

Topchy, A., Jain, A. K., and Punch, W. (2004), "A Mixture Model for Cluster Ensembles," *Proceedings of the Fourth SIAM International Conference 2004 (SDM 2004)*.

Topchy, T., Minaei-Bidgoli, B., Jain, A. K., and Punch, W. F. (2004), "Adaptive Cluster Ensembles," *ICPR 2004*, 272-275.

Torgerson, W. S. (1958), *Theory and Methods of Scaling*, New York: Wiley.

Weingessel, A., Dimitriadou, E., and Hornik, K. (2001), "An Ensemble Method for Clustering," *In Proceedings of the International Conference on Artificial Neural Networks, Vienna (2001)*, 217-224.

Wolberg, W. H., and Mangasrian, O. L. (1990), "Multisurface method of pattern separation for medical diagnosis applied to breast cytology," *Applied Mathematics*, 87, 9193-9196.

Yan Yu, and Diane Lambert. (1999), "Fitting Trees to Functional Data With an Application to Time-of-Day Patterns," *Journal of Computation and Graphical Statistics*, 8, 749-762.

Yeung, K. Y., Haynor, D. R., and Ruzzo, W. L. (2001), "Validating clustering for gene expression data," *Bioinformatics*, 17, 309-318.

Young, F. W. (1981), "Quantitative Analysis of Qualitative Data," *Psychometrika*, 46, 357-388.

Zhang, H. (1998), "Classification Trees for Multiple Binary Responses," *Journal of the American Statistical Association*, 98, 180-193.