2.6. SPATIAL AUTO-CORRELATION

A major consideration in determining the density of sampling in seabed surveys is the distance over which the seabed assemblages remain homogeneous, within any defined habitat type. While it is inefficient to place sample sites too close together, information on seabed patterns will be inadequately sampled if sites are too far apart. Further, the same process means that spatial prediction of assemblages cannot be applied reliably at great distances from sampled sites, even if physical environment variables are measured and taken into account. This process may be quantified by several indices; one is the spatial auto-correlation distance.

Analyses of spatial auto-correlation distance were conducted to establish over which geographic distance the similarity between the species composition per site is maintained to some degree. This distance then would give an indication of the minimum inter-sample distance needed to represent the biota adequately, as well as, the maximum distance over which prediction would be feasible. Two measures were use to estimate the order of the auto-correlation distances for the Torres Strait datasets.

The semi-variogram was used for seabed physical data, such as depth and sediment attributes, and for the biological data, such as Bio_Code and seagrass presence/absence and algae p/a the Bray-Curtis dissimilarity metric was used as a function of the inter-site distances. The between station Bray-Curtis dissimilarities were plotted against the geographic distance between the corresponding stations. If correlation had been used as the metric, such plots are called ‘correlograms’ — with the Bray-Curtis dissimilarity metric, these plots may be termed ‘Bray-Curtis-ogram’s.

2.6.1. Physical Co-Variates

Most of the semi-variograms of the seabed physical data, (percentage mud, sand, gravel and rock, degree of grain sorting and grain size) indicated that the similarity between stations decreased rapidly as the distance between the stations increased. Beyond an inter-station distance of between about 0.05 and 0.1 decimal degrees (approx 6–11 km) any consistency between the stations had degenerated (Figure 2.6-1). However, for sediment carbonate content and depth the drop in similarity between stations with increased distance was more linear.

2.6.2. Biological Survey Data

The Bray-Curtis-ograms of the low-level and medium level biological survey datasets indicates that the dissimilarity between stations increased rapidly the further apart stations were (Figure 2.6-2 and Figure 2.6-3) and after a distance greater than about 0.05–0.1 decimal degrees (about 5–11 km) the coherence is lost. Although directional relationships were not investigated, it is possible that this may be slightly better for some directions and worse for others. This distance needs to be considered in the context of the whole region. Stratification of the region based on other (physico-chemical) co-variates can lead to better local results, as was observed in a similar nearby area of the Great Barrier Reef where stratified species assemblage data did have Bray-Curtis ranges of up to perhaps 20 km within physical strata (Pitcher et al 2002).

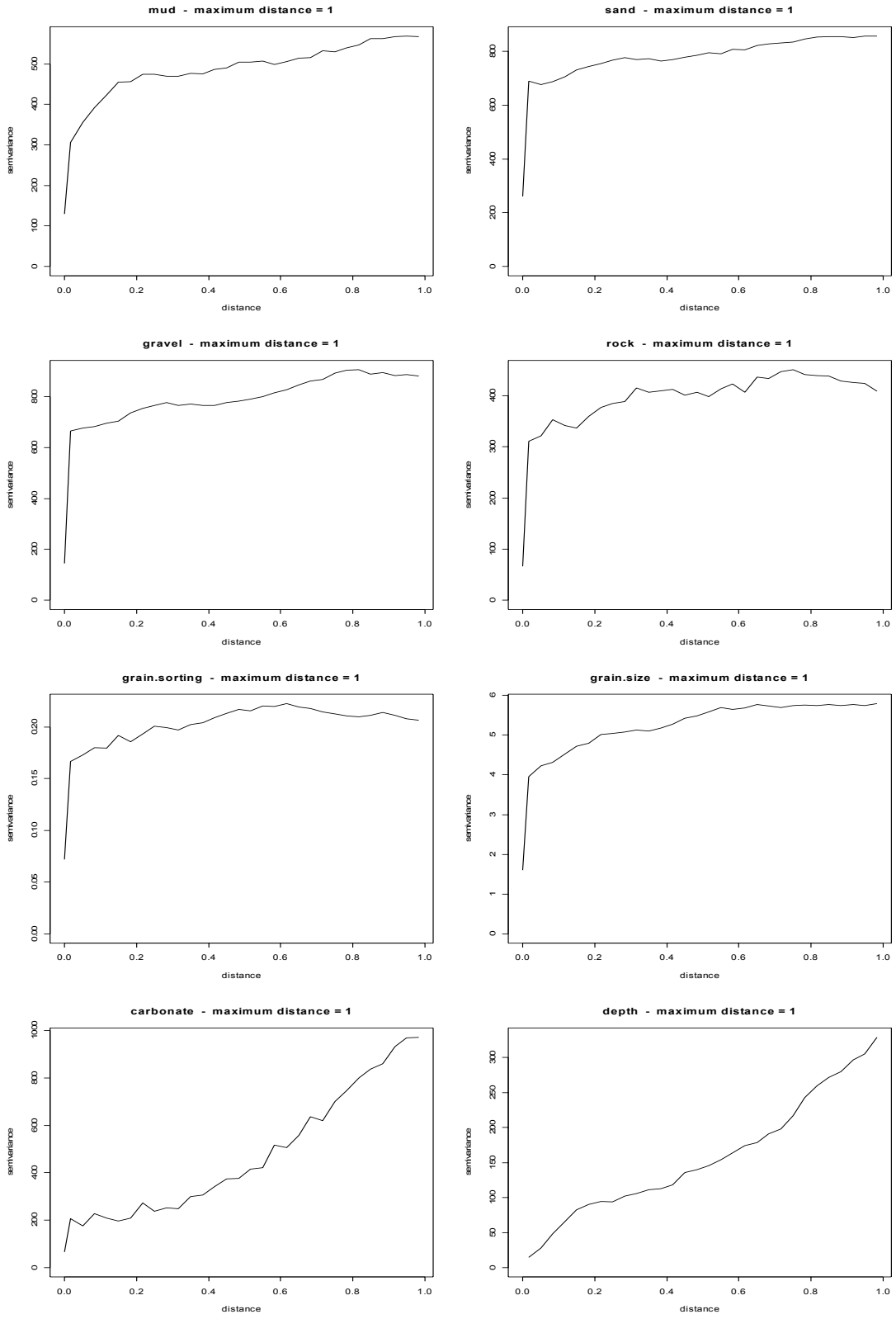


Figure 2.6-1 Semivariograms of the seabed physical data (percent of mud, sand, gravel and rock, degree of sorting, grain size, carbonate and depth) showing spatial autocorrelation between the survey sites, with distance (in decimal degrees).

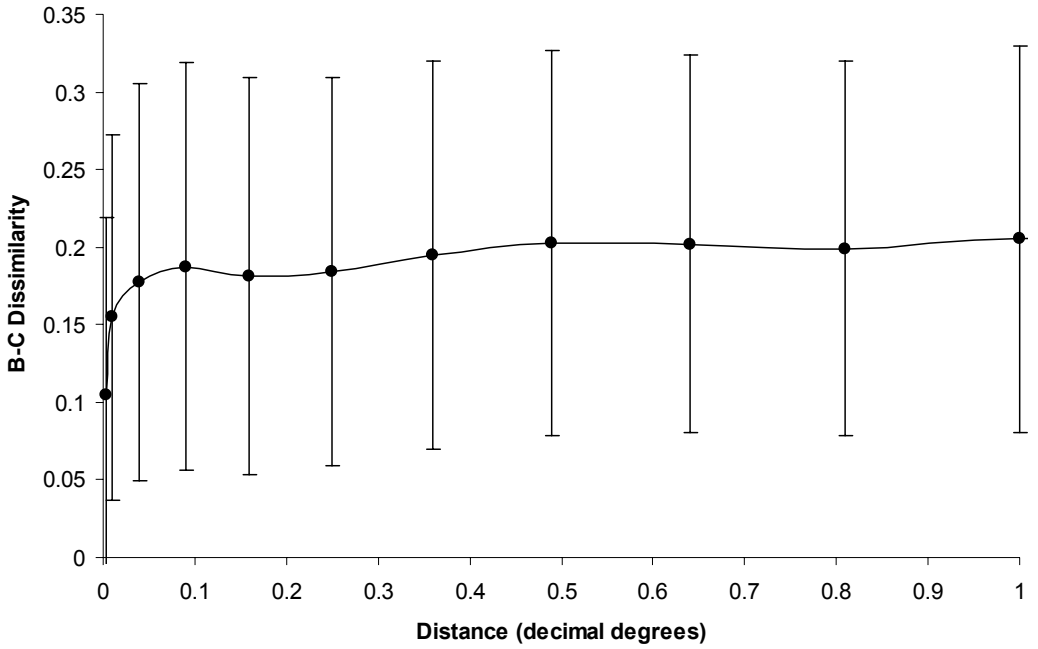


Figure 2.6-2 Bray-Curtis-ogram of the 'low-level' dataset (Table 2.4-1A), showing spatial autocorrelation for biological survey sites.

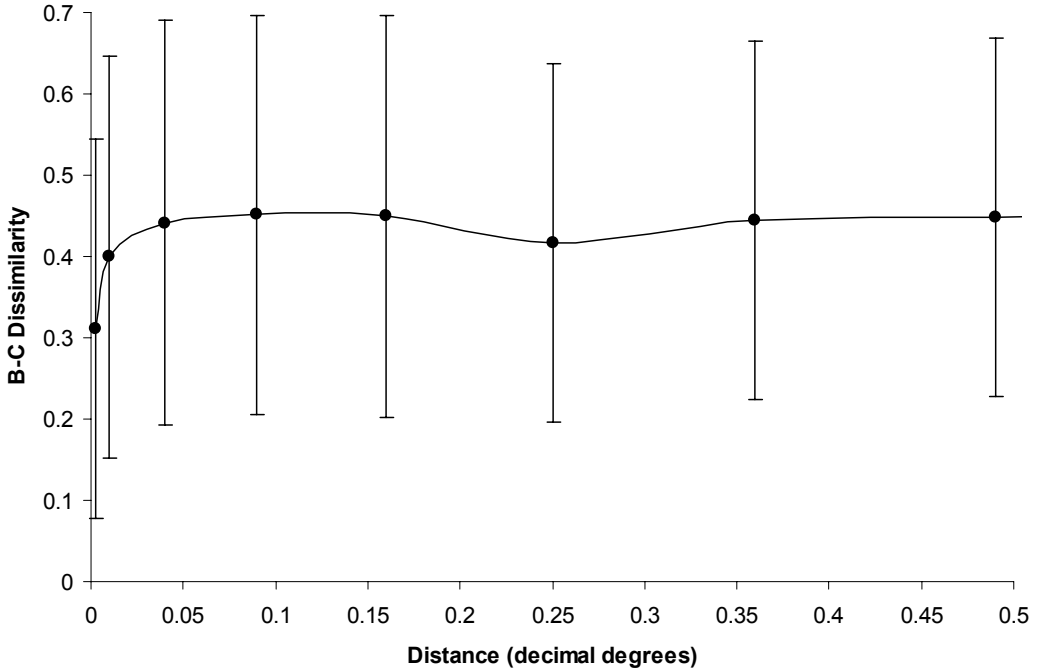


Figure 2.6-3 Bray-Curtis-ogram of the 'medium-level' dataset (Table 2.4-1B), showing spatial autocorrelation for biological survey sites. Note different x-axis scale cf Figure 2.6-2.

In summary, it is clear that the omnidirectional large scale distance that still carries some level of auto-correlation is in the order of ~5 km generally, and even within habitat strata, the spatial auto-correlation distance may extend to only ~10 – 20 km maximum. Consequently, if we do not have reliable co-variates with proven predictive capacity and relatively large homogeneous (geographic) areas, the inter-sample distance would need to be in the order of 10 km to enable some dependable spatial prediction capability. Spatial prediction mapping within these distance ranges is credible, but mapping seabed assemblages by extrapolating far beyond these distances would be highly unreliable and uncertain.

2.7. STATISTICAL CHARACTERISATION & BIO-PHYSICAL MODELING

Statistical approaches (including clustering) were used to characterise the mixtures of habitat facies in the biological survey datasets, and the relationships between these facies or clusters and the collated physical covariates was also examined. Based on any biophysical relationships and the physical covariate values at the ~50¸000 grid cells, the predicted cluster membership of each grid cell was mapped for a full coverage of Torres Strait. The biophysical relationships also indicate the relative importance of each physical covariate with respect to patterns in the biological data and provide weightings for each covariate in developing the stratification, which may be regarded as a biophysical characterisation of Torres Strait. These methods follow a similar approach as in Pitcher et al (2002).

In the case of the 'low-level' bio-survey dataset (Table 2.4-1A), a simple biophysical model was developed separately for each attribute (Sub_Code, Bio_Code, and Seagrass and Algae presence/absence), using the statistical method Linear Discriminant Analysis (LDA), which provides the best set of linear functions on the physical covariates that allocate the survey sites to the specified categories of the bio-survey attributes. The reliability of these functions (uncertainty of prediction) was also estimated by examining jack-knifed cross-validated error rates. Cross validation treats n-1 out of n training observations as a training set. It determines the discriminant functions based on these n-1 observations and then applies them to classify the one observation left out. This is done for each of the n training observations. The misclassification rate for each group is the proportion of sample observations in that group that are misclassified. This method achieves a nearly unbiased estimate but with a relatively large variance. The LDA functions were then applied to the Torres Strait wide coverage of physical covariates to map the estimated distribution of the facies of each bio-survey attribute for Torres Strait. The LDA also identifies which of the physical covariates were of use in allocating the sites to categories. Additionally, a simple cluster analysis was applied to characterise the combined attributes of the 'low-level' bio-survey dataset, and LDA was again used to develop a simple biophysical model for the clusters, provide cross-validated error rates, and to map these for an estimate of the full coverage for Torres Strait.

In the case of the 'medium-level' dataset (Table 2.4-1B), cluster analysis was applied to characterise the mixtures of these attributes, and the relationship between these clusters and the collated physical covariates, and their uncertainty, was examined and mapped using LDA, as above.

A similar approach was followed with the Torres Strait effects of trawling dataset that included detailed species-biomass distribution and abundance information from samples of seabed fish from

research trawls. Cluster analysis was applied to these data to provide a number of relatively homogenous mixtures of fish species, and, as above, LDA was used to develop biophysical models and map the fish assemblages to characterise the vicinity of the sampling.

2.7.1. 'Low-level' bio-survey dataset

2.7.1.1. Substrate Classification (SUB_CODE):

The substrate classification (SUB_CODE) data, widely available from Torres Strait seabed surveys, was defined as:

1 = >50% hard substratum,
2 = 10-50% hard substratum,
3 = rubble (<10% hard & >15% rubble),
4 = sand (muddy sand & sand),
5 = mud (mud & sandy mud)

The results from forward stepwise LDA on SUB_CODE categories alone showed the following. Based on the physical covariate data that was associated with the survey sites, categories 1 & 2, and 2 & 3 were most similar and categories 4 & 5 were quite different to the others, with category 3 intermediate (Table 2.7-1). The most important covariates for discriminating SUB_CODE categories were seabed current stress, CARS Silicate standard deviation, CARS Salinity average, and CARS Oxygen sd (Table 2.7-2). Sediment covariates were moderately important. The overall jack-knifed discrimination accuracy was 51% correct and ranged from 27% correct for category 2 to 80% correct for category 5 (Table 2.7-3) — in most cases the LDA confuses the most similar (and adjacent) category, so performs somewhat better than the raw error rates indicate.

The mapping of predicted SUB_CODE over Torres Strait (Figure 2.7-1) closely reflects the available data (Figure 2.4-2) and can be considered a reasonably reliable representation of the distribution of the gross substratum types.

Table 2.7-1. Between groups F-matrix for the physical covariate data associated with the SUB_CODE categories at sites (df = 9, 836). Small F-values indicate that the covariates of respective categories are more similar; large values indicate that the covariates of categories are more different.

	SUB_CODE Category				
	1	2	3	4	5
1	0.000				
2	3.800	0.000			
3	9.071	7.443	0.000		
4	21.225	25.293	23.337	0.000	
5	36.898	43.053	46.391	25.475	0.000

Table 2.7-2. F-values for physical covariates assessed by LDA for SUB_CODE. Increasing F-values indicate covariates that were more important for discriminating SUB_CODE categories. Covariates in the left columns were included in the final model; those on the right were excluded.

Covariate included in model		**Covariate excluded from model**	
Variable	F-value	Variable	F-value
M_BSTRESS	69.26	SW_CHLA_SD	3.39
CARS_SI_SD	21.18	DEM_SLOPE	2.99
CARS_S_AV	12.03	AGSO_DEM	2.68
CARS_O2_SD	11.23	SW_D_B_IRRAD	2.39
auDB_ROCK	8.29	SW_K490_SD	2.30
auDB_MUD	7.75	SW_CHLA_Y_AV	2.28
CARS_PO4_S	5.24	SW_K490_Y_AV	1.99
auDB_GRNSZ	4.65	DEM_ASPEC	1.55
auDB_GRAVEL	4.49	CARS_T_SD	1.55
		auDB_SAND	1.13
		auDB_CRBNT	1.12
		SW_K490_Y_SD	1.12
		CARS_NO3_S	0.97
		CARS_S_SD	0.94
		CARS_O2_AV	0.68
		SW_CHLA_Y_SD	0.68
		auDB_GRNSRT	0.53
		CARS_NO3_A	0.46
		CARS_SI_AV	0.46
		CARS_T_AV	0.45
		CARS_PO4_A	0.17

Table 2.7-3. Jackknifed classification matrix for discriminating SUB_CODE categories on physical covariates. Cases in row categories classified into columns. % correct shows number of cases in each row classified into the correct column. Diagnostic statistics: Wilks' lambda=0.439, approx.F=21.421, df=36, 3134.

	SUB_CODE Category					
	1	**2**	**3**	**4**	**5**	**%correct**
1	29	12	4	2	3	**58**
2	45	43	32	25	12	**27**
3	35	27	79	43	3	**42**
4	20	25	45	202	63	**57**
5	0	2	3	15	80	**80**
Total	**129**	**109**	**163**	**287**	**161**	**51**

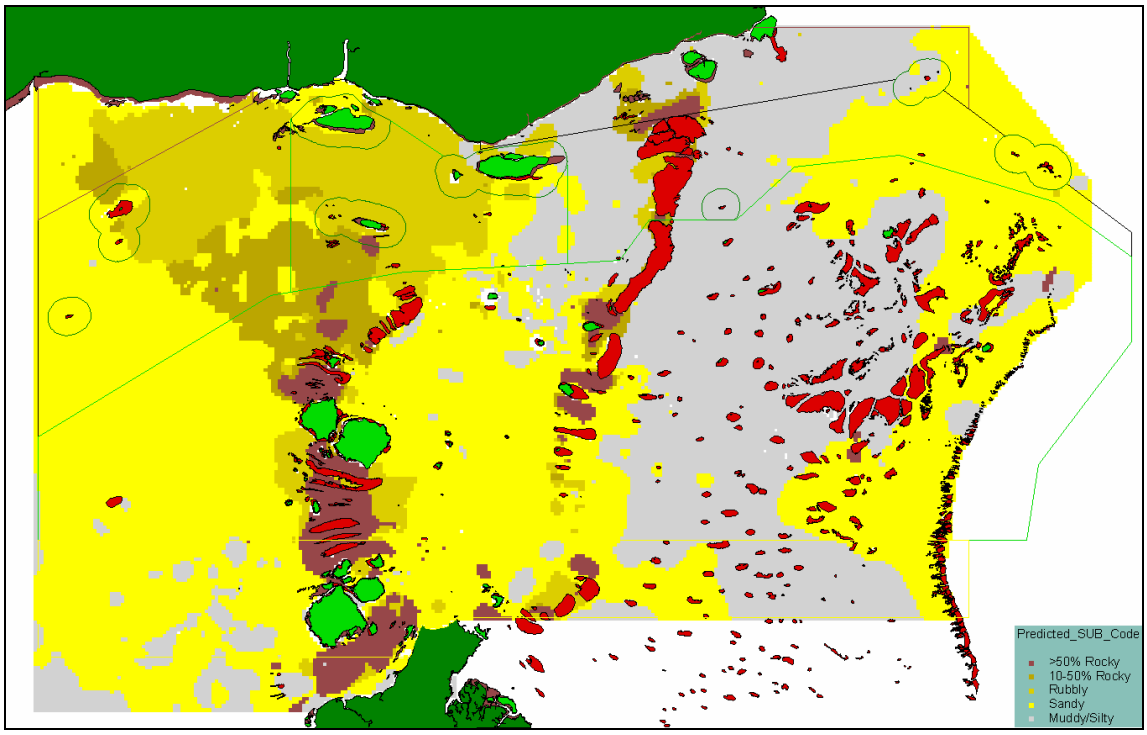


Figure 2.7-1. Mapping of predicted SUB_CODE over Torres Strait based on LDA functions applied to the gridded physical covariate dataset.

2.7.1.2. Epibenthos Classification (BIO_CODE):

The epibenthos classification (BIO_CODE), widely available from Torres Strait seabed surveys, was defined as (see section 2.1.1 for more details):

1 = dense fauna,
2 = sparse fauna,
3 = very sparse fauna,
4 = no fauna

The results from forward stepwise LDA on BIO_CODE categories alone showed the following. Based on the physical covariate data that was associated with the survey sites, categories 1 & 2, and 2 & 3 were most similar and category 4 was quite different to the others (Table 2.7-4). The most important covariates for discriminating BIO_CODE categories were seabed current stress, SeaWiFS average annual attenuation at 490 nm (turbidity), CARS nitrate standard deviation, and sediment gravel fraction (Table 2.7-5). The overall jack-knifed discrimination accuracy was 44% correct and ranged from 33% correct for category 2 to 67% correct for category 4 (Table 2.7-6) — in most cases the LDA confuses the most similar (and adjacent) category, so performs somewhat better than the raw error rates indicate.

The mapping of predicted BIO_CODE over Torres Strait (Figure 2.7-2) closely reflects the available data (Figure 2.4-2) and can be considered a reasonably reliable representation of the distribution of the gross epibenthos facies. Nevertheless, it is unlikely that the area in the vicinity of the outer barrier has the predicted extensive gardens of dense epibenthos.

Table 2.7-4. Between groups F-matrix for the physical covariate data associated with the BIO_CODE categories at sites (df = 6, 840). Small F-values indicate that the covariates of respective categories are more similar; large values indicate that the covariates of categories are more different.

	BIO_CODE Category			
	1	2	3	4
1	0.000			
2	5.395	0.000		
3	15.777	6.259	0.000	
4	42.306	27.732	32.097	0.000

Table 2.7-5. F-values for physical covariates assessed by LDA for BIO_CODE. Increasing F-values indicate covariates that were more important for discriminating BIO_CODE categories. Covariates in the left columns were included in the final model; those on the right were excluded.

Covariate included in model		**Covariate excluded from model**	
Variable	F-value	Variable	F-value
M_BSTRESS	31.82	auDB_CRBNT	3.08
SW_K490_Y_AV	10.99	DEM_SLOPE	2.64
CARS_NO3_SD	10.31	CARS_O2_AV	2.6
auDB_GRAVEL	10.13	SW_CHLA_SD	2.46
auDB_ROCK	6.11	CARS_T_AV	2.28
CARS_NO3_AV	4.06	CARS_PO4_AV	2.17
		SW_K490_Y_SD	2.16
		SW_K490_SD	1.8
		CARS_T_SD	1.73
		CARS_S_AV	1.46
		SW_CHLA_Y_SD	1.44
		CARS_O2_SD	1.37
		auDB_GRNSZ	1.16
		CARS_S_SD	1.04
		auDB_GRNSRT	0.99
		CARS_SI_SD	0.79
		SW_CHLA_Y_AV	0.46
		CARS_PO4_S	0.35
		CARS_SI_AV	0.2
		AGSO_DEM	0.16
		DEM_ASPEC	0.13
		auDB_SAND	0.12
		auDB_MUD	0.06
		SW_D_B_IRRAD	0.06

Table 2.7-6. Jackknifed classification matrix for discriminating BIO_CODE categories on physical covariates. Cases in row categories classified into columns. % correct shows number of cases in each row classified into the correct column. Diagnostic statistics: Wilks' lambda=0.670, approx.F= 20.114, df= 18, 2376.

	SUB_CODE Category				
	1	**2**	**3**	**4**	**%correct**
1	44	28	13	5	**49**
2	37	42	31	17	**33**
3	96	76	146	107	**34**
4	14	10	45	138	**67**
Total	**191**	**156**	**235**	**267**	**44**

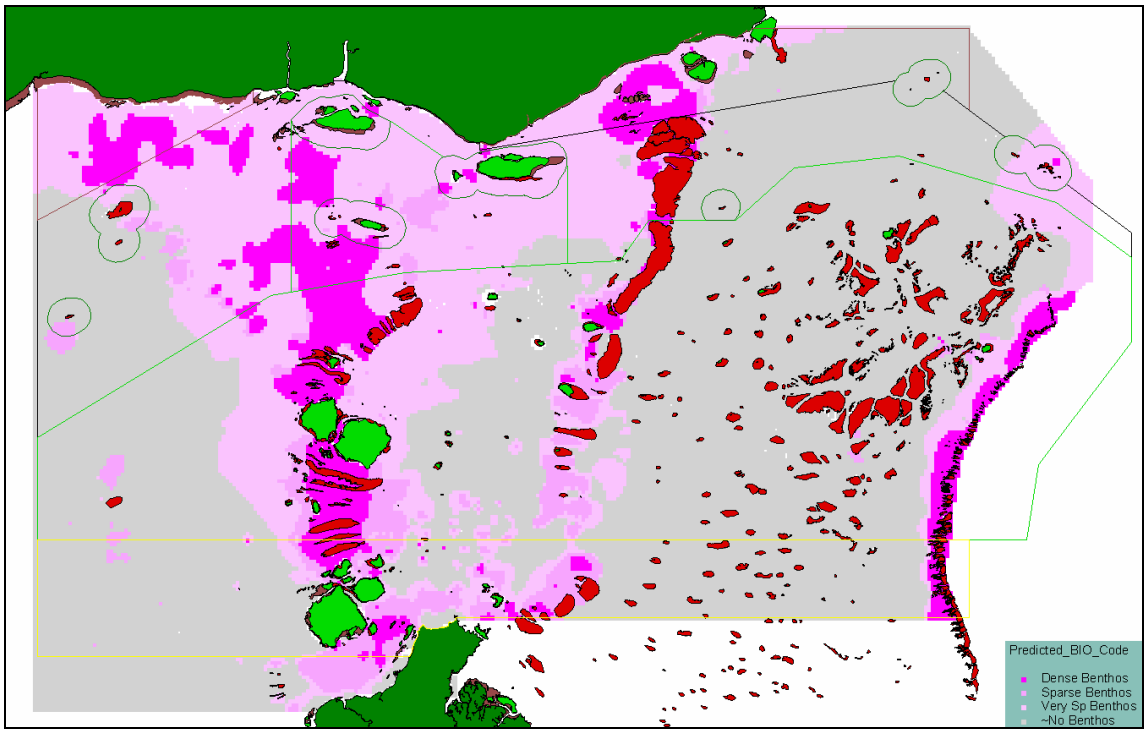


Figure 2.7-2. Mapping of predicted epibenthos BIO_CODE over Torres Strait based on LDA functions applied to the gridded physical covariate dataset.

2.7.1.3. Algae Presence/Absence (ALGAE_PA):

The Algae presence/absence data was widely available from Torres Strait seabed surveys, and was defined as: 0= absent, 1= present. The results from forward stepwise LDA on Algae_PA alone showed the following. The physical covariate data that was associated with the survey sites for presence and absence was quite different (F=22.847, df=5, 843). The most important covariates for discriminating Algae_PA were CARS Oxygen average, CARS Salinity average (Table 2.7-7). The overall jack-knifed discrimination accuracy was 73% correct and was similar for both presence and absence (Table 2.7-8).

The mapping of predicted Algae_PA over Torres Strait (Figure 2.7-3) closely reflects the available data (Figure 2.4-5), indicating that the numerous algal species are widely distributed through much of Torres Strait, and can be considered a reasonably reliable representation of the distribution of the gross algal distribution.

Table 2.7-7. F-values for physical covariates assessed by LDA for ALGAE_PA. Increasing F-values indicate covariates that were more important for discriminating ALGAE_PA categories. Covariates in the left columns were included in the final model; those on the right were excluded.

Covariate included in model		Covariate excluded from model	
Variable	F-value	Variable	F-value
CARS_O2_AV	69.53	CARS_PO4_S	3.13
CARS_S_AV	29.68	AGSO_DEM	1.24
auDB_ROCK	6.04	CARS_PO4_AV	1.14
CARS_NO3_AV	5.82	auDB_GRNSZ	0.91
auDB_CRBNT	5.26	SW_CHLA_Y_AV	0.76
		SW_CHLA_SD	0.76
		SW_CHLA_Y_SD	0.71
		auDB_MUD	0.64
		CARS_T_AV	0.61
		CARS_SI_AV	0.42
		M_BSTRESS	0.41
		SW_K490_Y_AV	0.36
		CARS_S_SD	0.32
		CARS_T_SD	0.25
		auDB_GRAVEL	0.18
		SW_D_B_IRRAD	0.18
		SW_K490_Y_SD	0.17
		DEM_SLOPE	0.13
		CARS_NO3_S	0.08
		CARS_O2_SD	0.06
		auDB_GRNSRT	0.06
		CARS_SI_SD	0.05
		auDB_SAND	0.02
		DEM_ASPEC	0.01
		SW_K490_SD	0.01

Table 2.7-8. Jackknifed classification matrix for discriminating ALGAE_PA categories on physical covariates. Cases in row categories classified into columns. % correct shows number of cases in each row classified into the correct column. Diagnostic statistics: Wilks' lambda=0.881, approx.F= 22.847, df= 5, 843.

	Algae P/A		
	0	**1**	**%correct**
0	57	22	**72**
1	205	565	**73**
Total	**262**	**587**	**73**

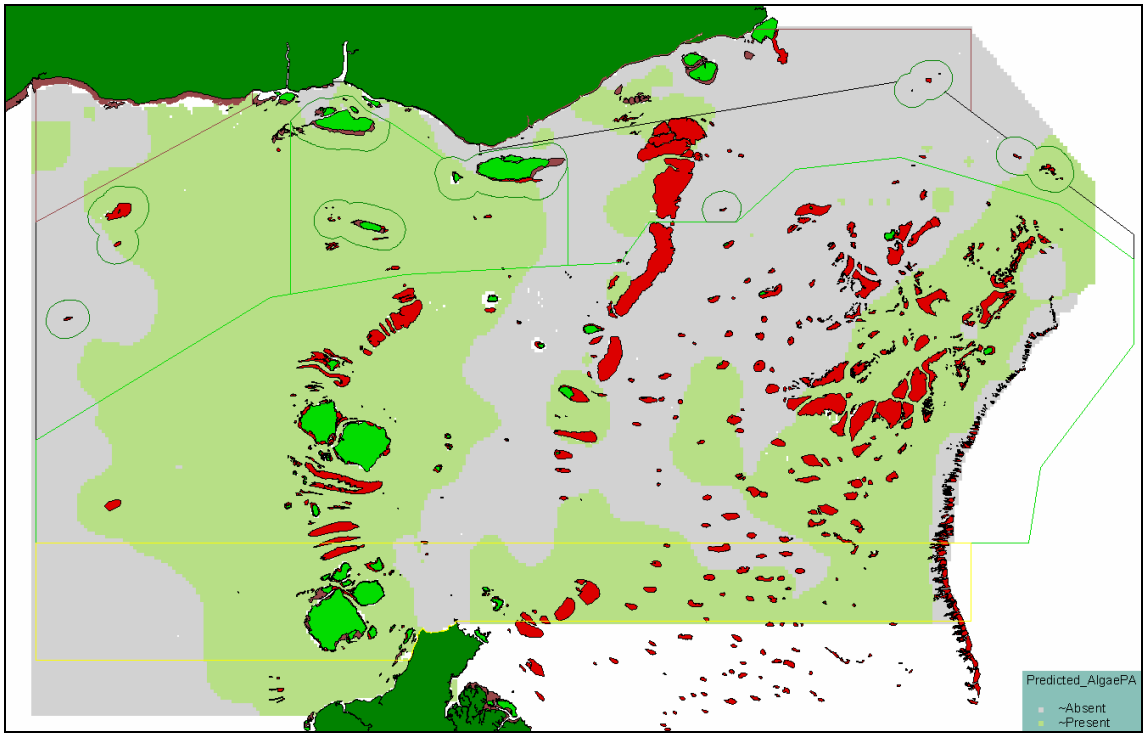


Figure 2.7-3. Mapping of predicted Algae P/A over Torres Strait based on LDA functions applied to the gridded physical covariate dataset.

2.7.1.4. Seagrass Presence/Absence (SEAGRASS_PA):

The Seagrass presence/absence data was widely available from Torres Strait seabed surveys, and was defined as: 0= absent, 1= present. The results from forward stepwise LDA on Seagrass_PA alone showed the following. The physical covariate data that was associated with the survey sites for presence and absence was quite different (F=28.132, df=6, 842). The most important covariates for discriminating Seagrass_PA were CARS Phosphate standard deviation, seabed current stress, CARS Nitrate standard deviation, sediment carbonate composition, and seabed irradiance estimated from SeaWiFS attenuation (Table 2.7-9). The overall jack-knifed discrimination accuracy was 68% correct and was slightly better for presence than absence (Table 2.7-10).

The mapping of predicted Seagrass_PA over Torres Strait (Figure 2.7-4) closely reflects the available data (Figure 2.4-5), indicating that the several seagrass species are mainly distributed through central western Torres Strait, and can be considered a reasonably reliable representation of the distribution of the gross seagrass distribution. Nevertheless, it is unlikely that the area in the vicinity of the outer barrier has the predicted seagrass presence.

Table 2.7-9. F-values for physical covariates assessed by LDA for SEAGRASS_PA. Increasing F-values indicate covariates that were more important for discriminating SEAGRASS_PA categories. Covariates in the left columns were included in the final model; those on the right were excluded.

Covariate included in model		**Covariate excluded from model**	
Variable	F-value	Variable	F-value
CARS_PO4_SD	31.55	SW_K490_SD	3.42
M_BSTRESS	29.04	CARS_T_AV	2.88
CARS_NO3_SD	25.46	SW_CHLA_SD	1.97
auDB_CRBNT	17.41	auDB_MUD	1.63
SW_D_B_IRRAD	12.48	CARS_T_SD	1.30
auDB_ROCK	5.89	auDB_GRNSRT	1.20
		SW_K490_Y_SD	1.17
		CARS_S_SD	1.00
		auDB_SAND	0.89
		SW_CHLA_Y_AV	0.77
		CARS_PO4_A	0.33
		CARS_SI_AV	0.21
		AGSO_DEM	0.19
		DEM_SLOPE	0.08
		auDB_GRAVEL	0.05
		SW_K490_Y_AV	0.04
		DEM_ASPEC	0.03
		CARS_SI_SD	0.03
		CARS_S_AV	0.03
		CARS_O2_AV	0.02
		CARS_NO3_A	0.01
		CARS_O2_SD	0.00
		auDB_GRNSZ	0.00
		SW_CHLA_Y_SD	0.00

Table 2.7-10. Jackknifed classification matrix for discriminating SEAGRASS_PA categories on physical covariates. Cases in row categories classified into columns. % correct shows number of cases in each row classified into the correct column. Diagnostic statistics: Wilks' lambda=0.833, approx.F= 28.132, df= 6, 842.

	Algae P/A		
	0	**1**	**%correct**
0	338	196	**63**
1	77	238	**76**
Total	**415**	**434**	**68**

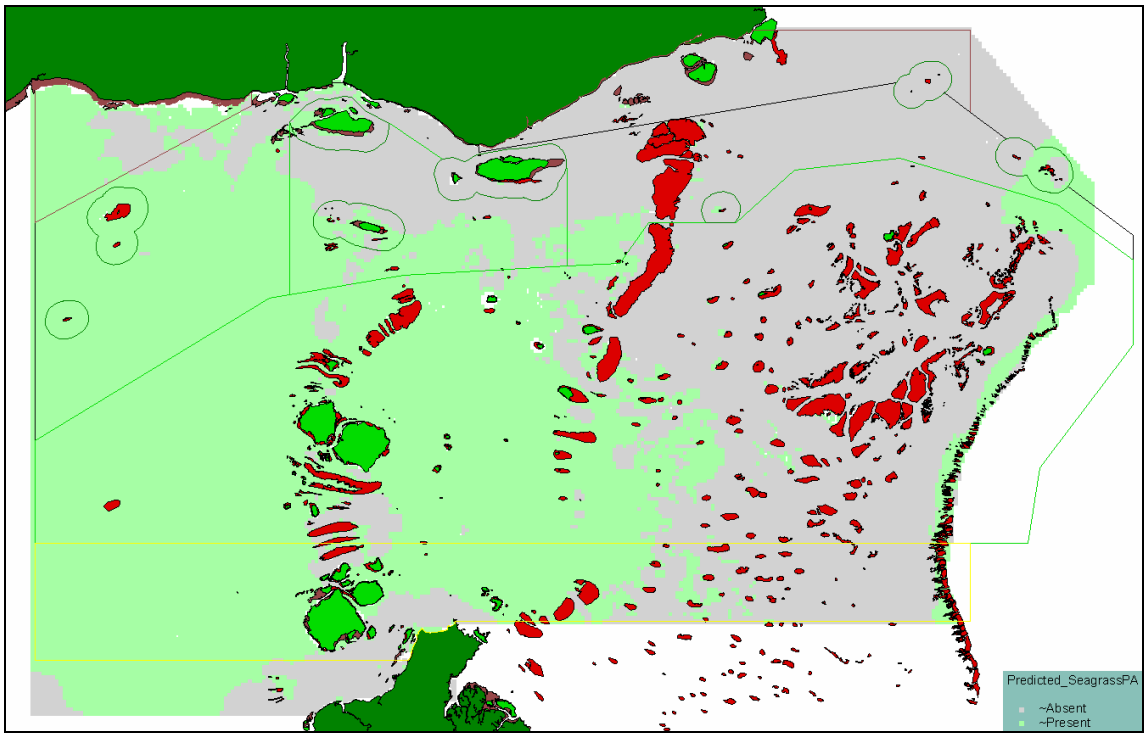


Figure 2.7-4. Mapping of predicted Seagrass P/A over Torres Strait based on LDA functions applied to the gridded physical covariate dataset.

2.7.1.5. Clustered Sub_Code, Bio_Code, Algae & Seagrass P/A:

The broad coverage 'low-level' bio-survey attributes were clustered to characterize the predominant mixtures of these habitat facies. Initially, the dataset was clustered into 4, 6 and 9 groups, using two algorithms for the K-means method (Euclidean & Sums-of-Squares), then the LDA jack-knifed classification performance of each was estimated to examine the trade-off between information detail (number of clusters) and potential mapping reliability (classification accuracy). In this case, 6 clusters appeared to be a reasonable compromise between information content and bio-physical classification success, and the Euclidean algorithm performed slightly better than SS.

From the statistics associated with each of the 6 clusters, it is possible to characterise them broadly as follows (Figure 2.7-5):

1: muddy/sandy, very sparse/no benthos, ~90% likelihood algae, ~40% likelihood seagrass
2: rubbly/some hard sub, dense/sparse benthos, ~95% likelihood algae, ~25% likelihood seagrass
3: sandy, sparse/very sparse benthos, ~65% likelihood algae, ~20% likelihood seagrass
4: rubble/some hard sub, very sparse benthos, ~95% likelihood algae, ~55% likelihood seagrass
5: mostly hard sub, dense/sparse benthos, ~100% likelihood algae, ~10% likelihood seagrass
6: some hard sub, very sparse benthos, ~95% likelihood algae, ~5% likelihood seagrass

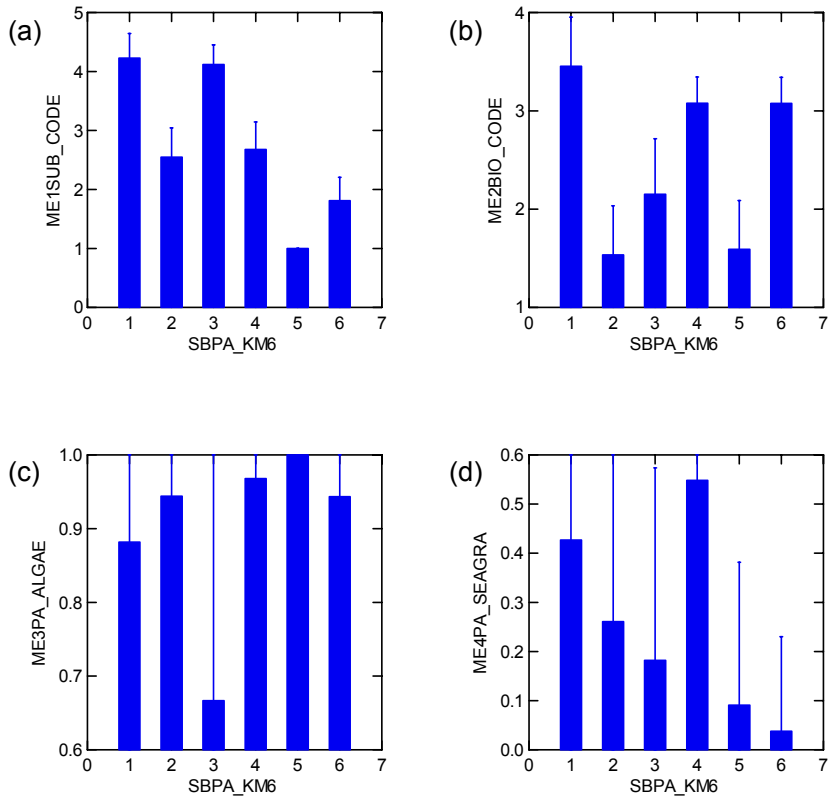


Figure 2.7-5. Mean and standard deviation of attributes for six K-means clusters of the 'low-level' habitat dataset: (a) seabed Sub_Code, (b) epibenthos Bio-Code, (c) Algae presence/absence, and (d) Seagrass presence/absence.

The results from forward stepwise LDA on the six Clusters showed the following. Based on the physical covariate data that was associated with the survey sites, cluster types 1 & 3, and 2 & 4 were most similar, with cluster 2 most different from 1 and clusters 4, 5, 6 intermediate from 1 (Table 2.7-11). The most important covariates for discriminating cluster types were seabed current stress, SeaWiFS estimated Chlorophyll a, with four other covariates moderately important (Table 2.7-12). The overall jack-knifed discrimination accuracy was 45% correct and ranged from 9% correct for cluster 3, which was fewest in number, to 65% correct for cluster 1, which was the most numerous type (Table 2.7-13).

The mapping of predicted cluster membership over Torres Strait (Figure 2.7-6) closely reflects the patterns for Sub_Code (Figure 2.7-1) and Bio_Code (Figure 2.7-2), which is not unexpected as these attributes dominated the cluster analysis. The map can be considered a reasonably reliable representation of the distribution of the broad habitat types in Torres Strait.

Table 2.7-11. Between groups F-matrix for the physical covariate data associated with the habitat cluster types at sites (df = 19, 853). Small F-values indicate that the covariates of respective categories are more similar; large values indicate that the covariates of categories are more different.

	Category					
	1	2	3	4	5	6
1	0					
2	46.916	0				
3	1.374	7.296	0			
4	33.746	2.555	5.184	0		
5	32.547	11.067	9.784	10.196	0	
6	23.899	5.506	5.919	4.864	4.117	0

Table 2.7-12. F-values for physical covariates assessed by LDA for habitat cluster types. Increasing F-values indicate covariates that were more important for discriminating cluster types. Covariates in the left columns were included in the final model; those on the right were excluded.

Covariate included in model		**Covariate excluded from model**	
Variable	F-value	Variable	F-value
M_BSTRESS	44.75	auDB_MUD	2.77
SW_CHLA_AV	14.56	CARS_T_SD	1.84
CARS_NO3_AV	8.10	auDB_ROCK	1.81
SW_K490_SD	7.62	CARS_SI_AV	1.32
auDB_GRNSZ	4.91	CARS_SI_SD	1.17
CARS_NO3_SD	4.06	CARS_S_SD	1.11
		auDB_CRBNT	0.98
		SW_K490_AV	0.96
		auDB_GRAVEL	0.92
		CARS_O2_AV	0.90
		CARS_O2_SD	0.89
		AGSO_DEM	0.88
		SW_CHLA_SD	0.88
		CARS_PO4_AV	0.79
		auDB_SAND	0.72
		CARS_T_AV	0.71
		CARS_S_AV	0.65
		CARS_PO4_SD	0.54
		auDB_GRNSRT	0.52

Table 2.7-13. Jackknifed classification matrix for discriminating habitat cluster types on physical covariates. Cases in row categories classified into columns. % correct shows number of cases in each row classified into the correct column. Diagnostic statistics: Wilks' lambda=0.570, approx.F=17.414, df=30, 3466.

	Cluster						
	1	2	3	4	5	6	%correct
1	278	36	36	51	16	14	65
2	18	43	17	32	19	32	27
3	18	3	3	2	2	5	9
4	18	33	17	35	18	34	23
5	2	1	5	6	20	10	45
6	4	11	7	6	11	14	26
Total	338	127	85	132	86	109	45

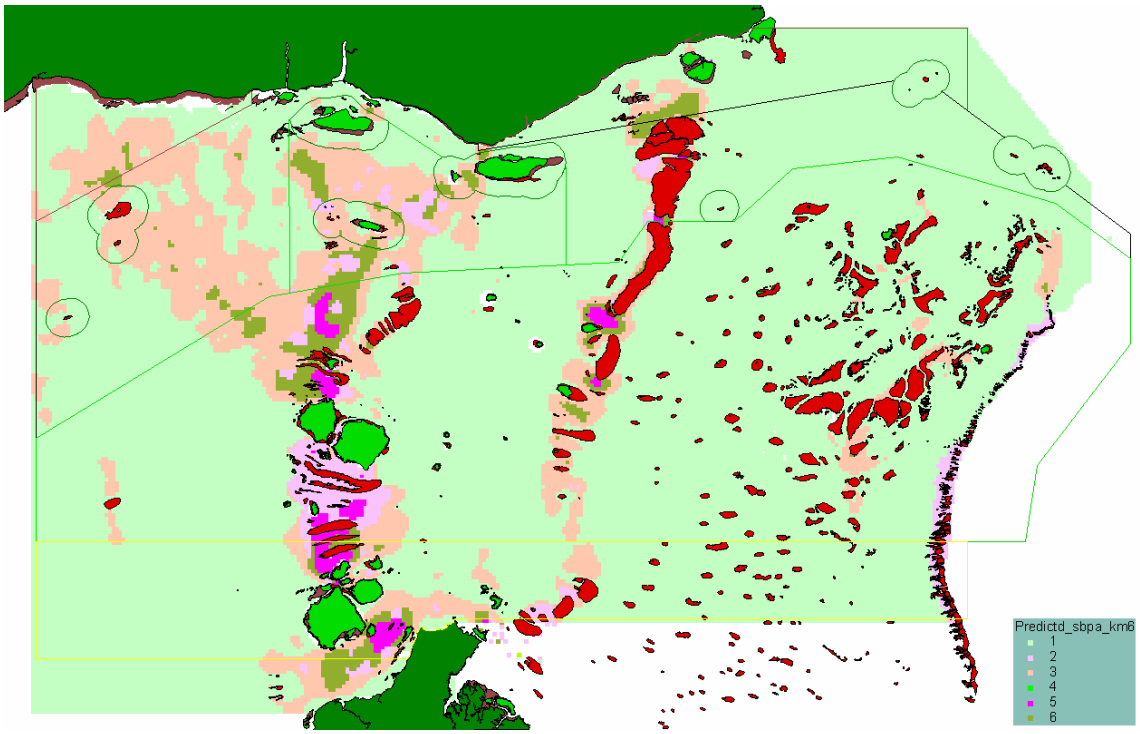


Figure 2.7-6. Mapping of predicted habitat cluster membership over Torres Strait based on LDA functions applied to the gridded physical covariate dataset. Habitat clusters were characterized as follows:

1: muddy/sandy, very sparse/no benthos, ~90% likelihood algae, ~40% likelihood seagrass
2: rubbly/some hard substrate, dense/sparse benthos, ~95% likelihood algae, ~25% likelihood seagrass
3: sandy, sparse/very sparse benthos, ~65% likelihood algae, ~20% likelihood seagrass
4: rubble/some hard substrate, very sparse benthos, ~95% likelihood algae, ~55% likelihood seagrass
5: mostly hard substrate, dense/sparse benthos, ~100% likelihood algae, ~10% likelihood seagrass
6: some hard substrate, very sparse benthos, ~95% likelihood algae, ~5% likelihood seagrass

2.7.2. 'Medium-level' bio-survey dataset

A slightly higher level of seabed habitat information was available at a reduced number of sites (Table 2.4-1B). The higher level biological survey data included the ordinal scale of epibenthos density (BIO_CODE), the percentage cover of seagrass and algae over the survey transects (PCT_TOT_SGRS and PCT_TOT_ALG respectively), and the estimated percentage cover of a number of sediment classes: mud-silt (PCT_MUD_SILT), sand-gravel (PCT_SAND_GRV), rubble (PCT_RUBBLE), consolidated rubble (PCT_CONS_RUB) and hard, rock pavement (PCT_SUB_HARD). Individually, these attributes were available from between 655 and 1196 sites in Torres Strait; however, there were only 421 sampling sites where records for all 8 of these attributes were recorded (Figure 2.7-7). These data were restricted to the central part of Torres Strait with most data collected from around the northern sections of the Warrior Reefs and around the island chain between Cape York and Papua New Guinea.

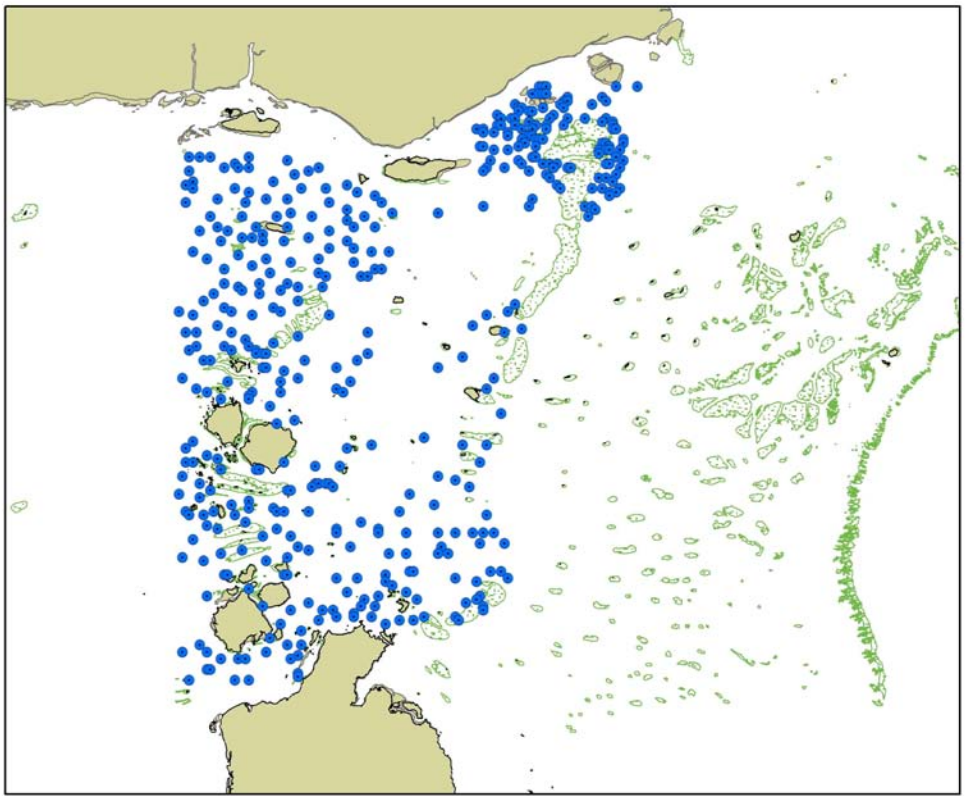


Figure 2.7-7. Map of Torres Strait showing the 421 sampling sites where records for all 8 "medium" level survey data were recorded.

In order to characterise the habitat facies represented by these data and examine the relationships between these data and the physical co-variate data, the habitat data were clustered and then linear discriminant function analysis was applied to determine how well the clusters could be described by the physical co-variate data. Only sites that had data for all 8 survey variables were included in these analyses.

The distribution of all survey variables was examined to determine whether transformation was necessary prior to analysis (Figure 2.7-8). BIO_CODE was originally recorded on an ordinal scale (see section 2.1.1 for more details). In an effort to prevent any particular variable having a disproportionate effect on the clustering process, we re-coded the values for BIO_CODE to match approximately the coverage given in the definitions (section 2.1.1) i.e. 1 = 60%, 2 = 30%, 3 = 5% and 4 = 0%. Similarly, the five sediment variables were rescaled to range between 0 and 20 rather than zero and 100. All variables were then $\log_e(x+1)$ transformed because of the highly right-skewed nature of most of their distributions (Figure 2.7-8).

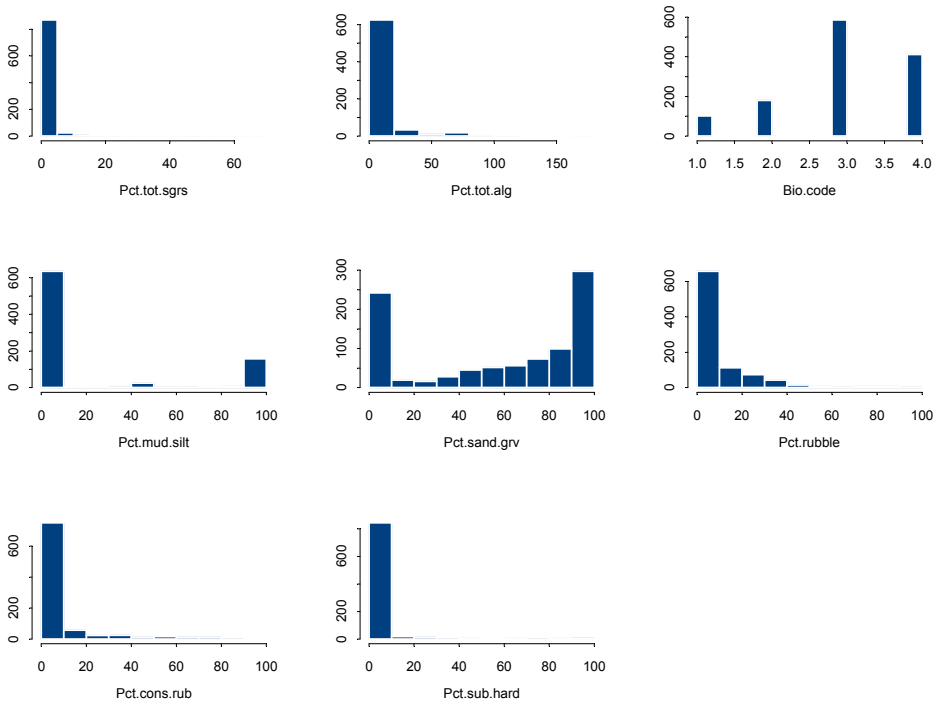


Figure 2.7-8. Histograms of the untransformed distributions of all "high" level survey data.

The Bray-Curtis dissimilarity distance metric was used to estimate the bio-survey data distance between all sites, then multidimensional scaling (MDS) with ordinal scaling in 4 dimensions was applied to reduce the dimensionality of the B-C matrix. The MDS of the "medium" level survey data revealed little in the way of discrete groupings. There was one central cloud of points containing the majority of observations surrounded by several disparate groupings containing several observations each (Figure 2.7-9).

The 4 dimensional MDS coordinates for each survey sites were then clustered using K-Means algorithm. As with the low-level dataset, a range of numbers of clusters was tried (4, 6 and 9), and as before, choosing an appropriate number of clusters was a compromise between the information content of the cluster characterisation and the biophysical classification success. The cluster membership of each site was then joined to the matching data for the collated physical co-variates (sections: 2.1.1 and 2.3) and a linear discriminant function analysis (LDA) was again used as above to develop a simple biophysical model for the clusters, provide cross-validated error rates, and to predict

and map the cluster membership of each 0.01 degree cell for an estimate of the full coverage for Torres Strait.

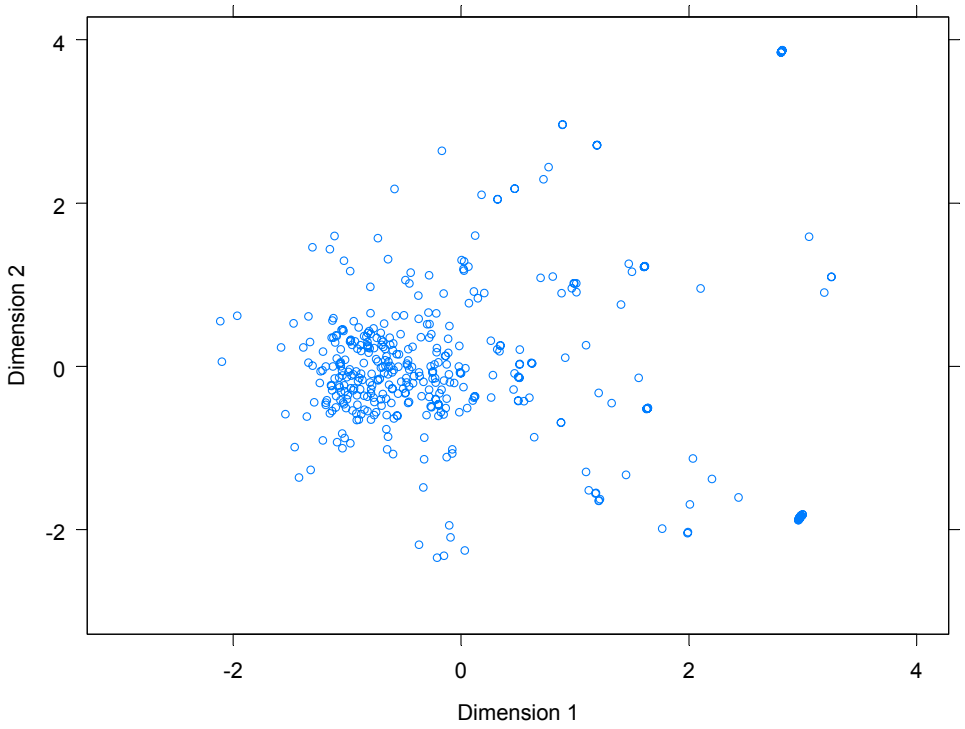


Figure 2.7-9. Ordination of the first and second dimensions of a multi dimensional scaling of the Bray-Curtis Dissimilarities of the "high" level Torres Strait survey data.

The results of the relative biophysical classification performance are shown in Table 2.7-14. The smaller the number of clusters, the better the performance of the discriminant functions in terms of correctly assigning a site to a cluster based on the values of the co-variates at that site, but the less information available about each cluster because their characteristics become more generalised as the total number of clusters decreases. In terms of the diagnostic statistics, 4 clusters gives the best performance (Table 2.7-14); however, 6 clusters were chosen to maximise the amount of biophysical information while retaining acceptable biophysical classification performance.

Table 2.7-14. Cluster diagnostics for K-Means clustering of the "high" level data.

	Wilks' Lambda	F value	Pr > F	Cross-validation: % correct
4 clusters	0.462	7.38	< 0.0001	51.2
6 clusters	0.384	4.98	< 0.0001	38.7
9 clusters	0.299	4.72	< 0.0001	22.2

Based on the statistics of the survey attributes (Figure 2.7-10), the composition of each cluster may be characterised as follows:

Cluster 1 – Sparse to dense epibenthos, sparse to medium algal cover, sandy with some rubble and consolidated rubble and pavement; no seagrass.

Cluster 2 – Sandy-muddy areas with little epibenthos, algae or seagrass.

Cluster 3 – Similar to cluster 1; sparse to dense epibenthos, sandy with some rubble and consolidated rubble. Some areas of pavement, although less than cluster 1. Less algal cover than cluster 1, very sparse seagrass cover.

Cluster 4 – Barren muddy areas having no epibenthos, algae or seagrass.

Cluster 5 – Very sparse epibenthos, sparse to medium cover of algae, sparse cover of seagrass. Generally sandy with some rubble, consolidated rubble and pavement.

Cluster 6 – Barren sandy areas with very little epibenthos and virtually no seagrass or algae.

Of the surveyed sites, the largest cluster was cluster 5 with 225 sites, cluster 4 was the smallest having only 12 sites, and clusters 1 2 3 and 6 were intermediate with between 39 to 50 sites.

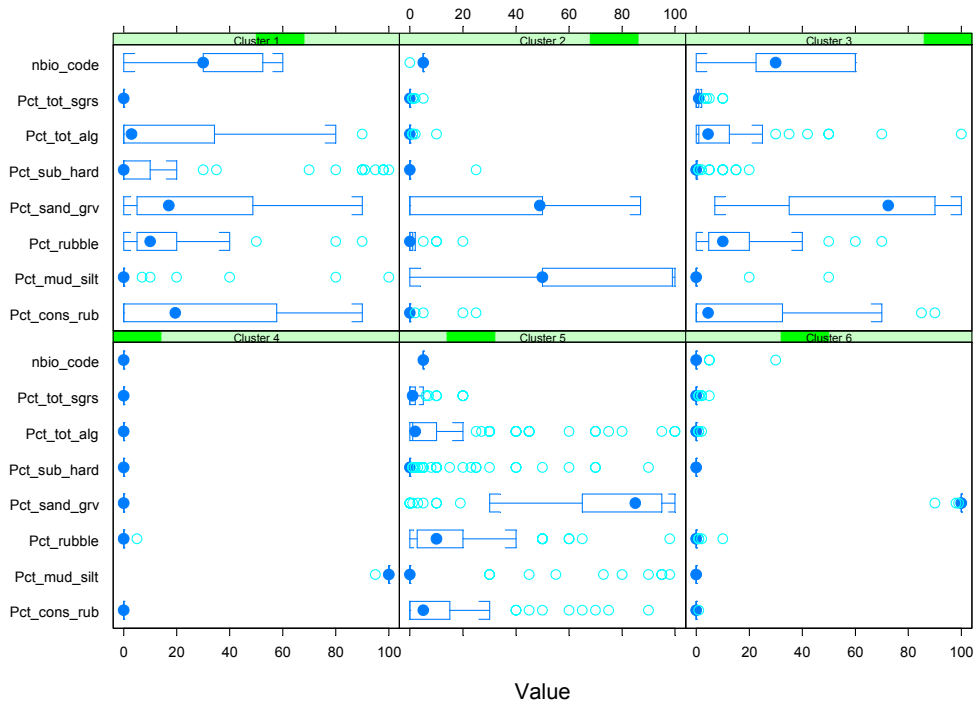


Figure 2.7-10. Box and whisker plots of distribution of the "high" level survey data among the 6 clusters. Median (dark blue closed circles), inter-quartile ranges (dark blue open boxes), 1.5 times the inter-quartile range (outer fences) except the outliers (light blue open circles).

The results of the LDA indicated that 17 of the physical co-variates appeared to be important in influencing the distribution of the clusters, and were included in the discriminant functions (Table 2.7-15). The two co-variates that had the highest partial r^2 and F values, and therefore were most

important in the discriminant functions were the modelled seabed current shear stress (M_BSTRESS) and the standard deviation of the annual SeaWiFS chlorophyll *a* estimate (SW_CHLA_Y_SD). The overall jack-knifed discrimination accuracy was ~39% correct and ranged from 0% correct for cluster 4, which had only 12 sites, to 56.4% correct for cluster 2 (Table 2.7-16).

Table 2.7-15. LDA statistics of co-variates included in the discriminant functions of the "medium" level survey data.

Co-Variate	Partial r^2	F-value	Pr > F
M_BSTRESS	0.1364	12.61	<.0001
SW_CHLA_Y_SD	0.0833	7.25	<.0001
CARS_O2_AV	0.0595	5.05	0.0002
CARS_S_AV	0.0557	4.71	0.0003
DEM_BATHY	0.0535	4.51	0.0005
CARS_PO4_S	0.0483	4.05	0.0014
CARS_O2_SD	0.0466	3.90	0.0018
auDB_ROCK	0.0448	3.74	0.0025
WTD_TRWL_E	0.0415	3.46	0.0045
SW_K490_Y_SD	0.0375	3.11	0.0091
auDB_GRAVEL	0.0353	2.92	0.0133
DEM_ASPECT	0.0307	2.52	0.0288
auDB_CRBNT	0.0305	2.51	0.0296
SW_CHLA_Y_AV	0.0292	2.40	0.0368
SW_K490_Y_AV	0.0281	2.30	0.0441
auDB_GRNSZ	0.0266	2.18	0.0560
CARS_PO4_AV	0.0218	1.78	0.1168

Table 2.7-16. Cross-validation summary for the performance of the discriminant functions in assigning survey sites to clusters. The number in the top of each cell is the number of sites; the number in the bottom is the percent of the total number of observations for that cluster. The shaded diagonal cells show the numbers of sites that were correctly assigned to each cluster (columns, from rows).

CLUSTER	1	2	3	4	5	6	Total
1	20 40.0	1 2.0	12 24.0	0 0.0	14 28.0	3 6.0	**50** **100.0**
2	2 5.1	22 56.4	3 7.7	0 0.0	5 2.8	7 18.0	**39** **100.0**
3	7 15.9	1 2.3	16 36.4	0 0.0	14 31.8	6 13.6	**44** **100.0**
4	1 8.3	4 33.3	1 8.3	0 0.0	2 16.7	4 33.3	**12** **100.0**
5	35 15.3	20 8.7	53 23.1	0 0.0	88 38.4	33 14.4	**229** **100.0**
6	1 2.1	7 14.9	10 21.3	0 0.0	12 25.5	17 36.2	**47** **100.0**
Total	**66** **15.7**	**55** **13.1**	**95** **22.6**	**0** **0.0**	**135** **32.1**	**70** **16.6**	**421** **100.0**

The discriminant functions were used to predict the cluster membership of every other 0.01 degree cell in the entire Torres Strait area, based on the values of the co-variates at each cell. This prediction was then mapped to show the distribution of predicted habitat clusters (Figure 2.7-11). The mapping suggests there are areas of sparse to dense epibenthos with some algal cover and a substrate consisting of sand and rubble (clusters 1 and 3) in the areas between many of the reefs and islands in the central Torres Strait. There are extensive sandy barren (cluster 6) areas in the south-west, central and eastern Torres Strait. Barren sandy-mud and muddy (clusters 2 and 4) areas extend across the northeast, to the east of the Warrior Reef complex and amongst the mid-shelf eastern reefs. The central area to the north of Cape York and parts of the north-western Torres Straits are characterised by a very sparse coverage of epibenthos, sparse algal and seagrass cover and a substrate consisting mainly of sand with patches of rubble, consolidated rubble and pavement (cluster 5).

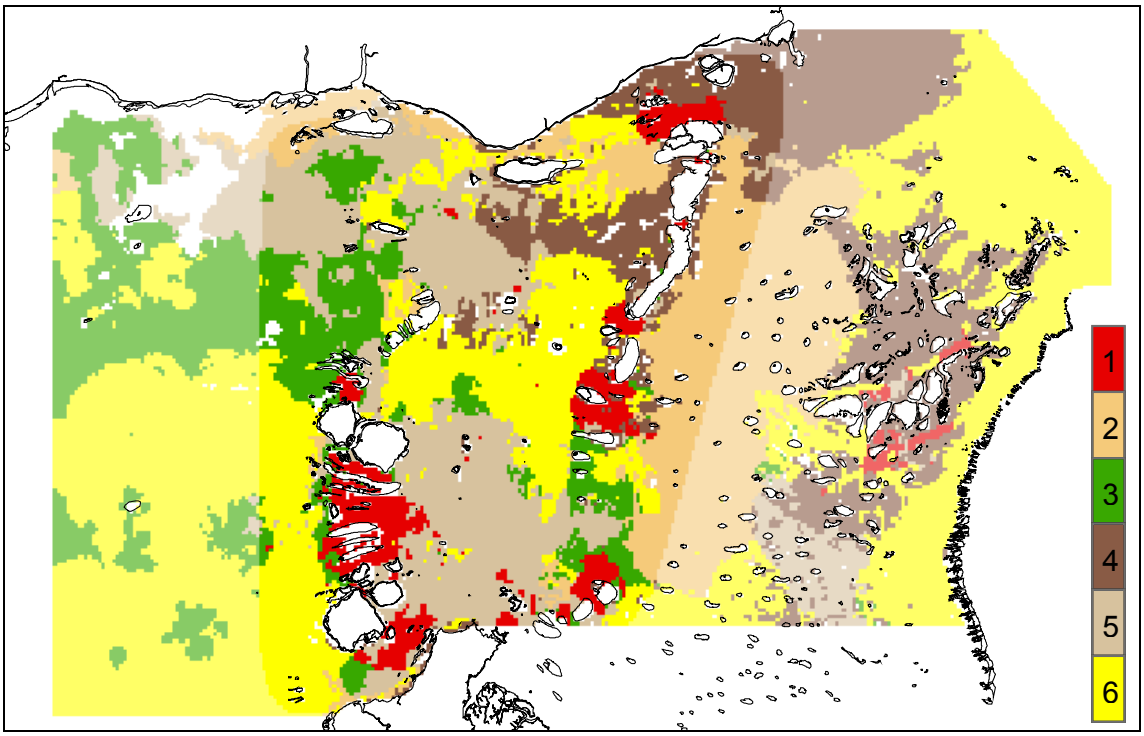


Figure 2.7-11. Mapping of predicted cluster membership (K-Means; 6 clusters) of the 'medium'-level survey data for the 0.01 degree gridded physical environment data, using linear discriminant function analysis. The darker shaded mapping represents the predictions within the area where survey data was collected; the lighter shaded mapping is outside the survey area.

2.7.3. Seabed Fish Dataset

The seabed fish dataset from the mid 1980's effects of trawling series of cruises, described in section 2.4.2.4 is characterised statistically here in a similar way as the seabed habitat data.

The standardised species sample weight data, averaged across all voyages, at each site was joined to the matching collated physical co-variate data. The mean catch rates were loge + 1 transformed and a

matrix of Bray-Curtis dissimilarity metrics was calculated. The matrix was then reduced by hierarchical agglomerative clustering (with group-average linking) and by MDS. The first 4 dimensions of the MDS were then clustered using the K-Means algorithm in a similar manner to the "medium" level survey data. Linear discriminant function analysis was then performed to determine how well the co-variate data could be used to predict and map the fish bycatch assemblage.

The cluster and MDS analyses of the mean catch rates showed fairly clear groupings based on whether the sample site was within the commercial trawl grounds or within a closure (Figure 2.7-12, Figure 2.7-13).

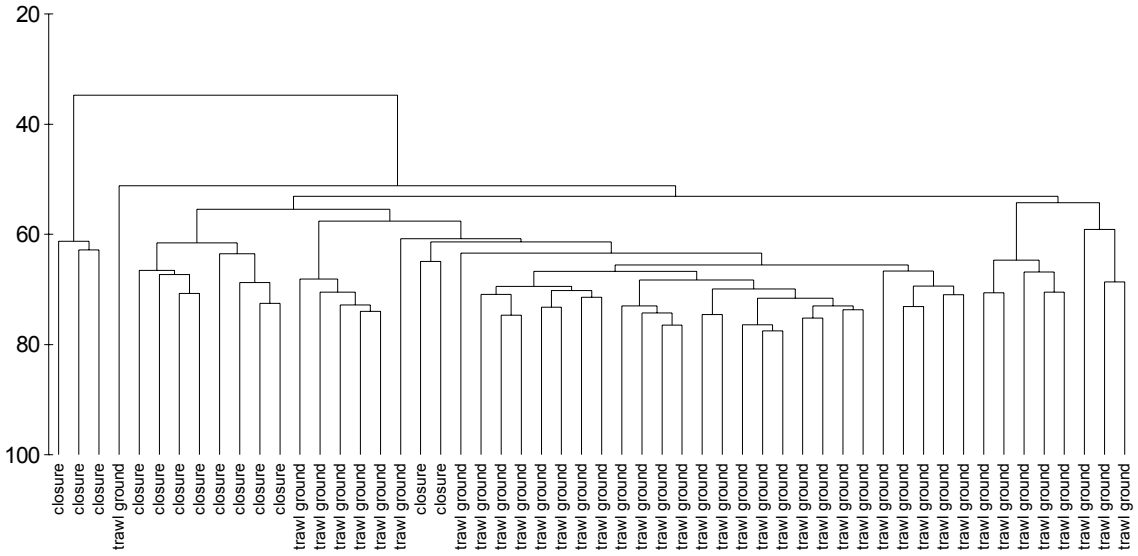


Figure 2.7-12. Hierarchical agglomerative clustering (group-average linking) of the mean catch rates of fish bycatch from prawn trawls done in the eastern Torres Strait during 1984-5. Each site is labeled based on whether it was located within the commercial trawl grounds or an area closed to fishing.

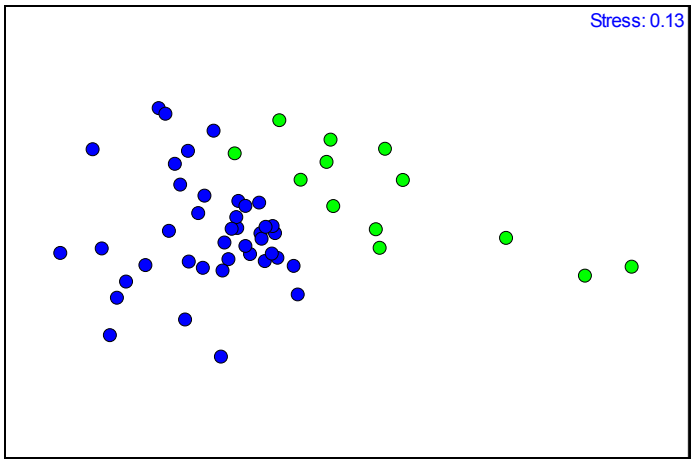


Figure 2.7-13. MDS ordination of the mean catch rates of fish bycatch from prawn trawl sampling conducted in the eastern Torres Strait during 1984-5. Each site is labelled based on whether it was located within the commercial trawl grounds (●) or an area closed to fishing (●).

Table 2.7-17. Mean catch rates of the 10 most abundant (biomass) fish in each cluster caught in the bycatch study in the central Torres Strait during the mid-1980s.

Cluster	Species	Family	Kg/ha	Habitat
1	*Lethrinus genivittatus*	Lethrinidae	7345.0	seagrass & weed beds
1	*Choerodon cephalotes*	Labridae	1815.6	coral reefs & nearby seagrass beds
1	*Pentapodus setosus*	Nemipteridae	1516.6	sand-rubble fringe of coral reefs
1	*Siganus canaliculatus*	Siganidae	1102.5	sand-weed areas
1	*Upeneus luzonius*	Mullidae	1066.0	Muddy bottoms
1	*Lethrinus laticaudis*	Lethrinidae	650.9	juveniles on seagrass; adults on coral reefs
1	*Gnathanodon speciosus*	Carangidae	572.3	coastal and offshore waters, usually near reefs
1	*Tragulichthys jaculiferus*	Diodontidae	507.6	soft bottom and trawl grounds
1	*Pseudomonacanthus elongatus*	Monacanthidae	471.7	?
1	*Psammoperca waigensis*	Centropomidae	428.7	rocky or coral reefs, frequently in weedy areas
2	Dasyatinae	Dasyatidae	3203.6	Sandy areas
2	*Priacanthus tayenus*	Priacanthidae	2389.3	coral reefs and rocky bottoms
2	*Saurida undosquamis*	Synodontidae	2057.4	sandy or muddy bottoms
2	*Nemipterus peroni*	Nemipteridae	1928.7	trawling grounds
2	*Saurida micropectoralis*	Synodontidae	1772.8	muddy bottoms
2	*Platycephalidae*	Platycephalidae	1549.0	sandy or muddy bottoms
2	*Nemipterus hexodon*	Nemipteridae	1264.0	trawling grounds
2	*Paramonacanthus japonicus*	Monacanthdae	967.1	weedy and sandy areas of coastal reefs
2	*Pomadsys maculatum*	Haemulidae	923.3	sandy or muddy bottoms
2	*Nemipterus furcosus*	Nemipteridae	883.3	trawling grounds
3	*Sphyrna mokarran*	Sphyrnidae	13034.8	coastal-pelagic, in passes and lagoons
3	*Stegastoma fasciatum*	Stegastomidae	4431.8	coastal & offshore waters in the vicinity of coral reefs
3	Dasyatinae	Dasyatidae	3910.4	sandy areas
3	*Rhina ancylostoma*	Rhyncobatidae	3798.7	coastal waters on mud or sand bottoms
3	*Nemipterus furcosus*	Nemipteridae	1540.8	trawling grounds
3	*Scolopsis taeniopterus*	Nemipteridae	1497.2	Sandy areas in the vicinity of coral reefs
3	*Priacanthus tayenus*	Priacanthidae	1336.7	coral reefs and rocky bottoms
3	*Myliobatus australis*	Myliobatidae	1266.2	?
3	*Nemipterus peroni*	Nemipteridae	1000.3	trawling grounds
3	*Pastinachus sephen*	Dasyatidae	977.6	flat sand or mud bottoms
4	*Saurida undosquamis*	Synodontidae	7307.5	sandy or muddy bottoms
4	*Carangoides talamparoides*	Carangidae	3895.8	coastal waters
4	*Leiognathus splendens*	Leiognathidae	3086.0	coastal waters, commonly on trawl grounds
4	*Priacanthus tayenus*	Priacanthidae	2946.3	coral reefs and rocky bottoms
4	*Paramonacanthus japonicus*	Monacanthdae	2923.1	weedy and sandy areas of coastal reefs
4	*Nemipterus peroni*	Nemipteridae	2239.5	trawling grounds
4	*Nemipterus furcosus*	Nemipteridae	2221.0	trawling grounds
4	*Scolopsis taeniopterus*	Nemipteridae	1938.6	sandy areas in the vicinity of coral reefs
4	*Leiognathus fasciatus*	Leiognathidae	1844.5	coastal waters, commonly on trawl grounds
4	*Saurida micropectoralis*	Synodontidae	1761.1	muddy bottoms

The K-Means clustering of the first 4 dimensions of the MDS ordination was restricted to 4 clusters because there was little evidence of finer clustering within the hierarchical dendrogram (Figure 2.7-12) of the only 54 sample sites in this dataset.

The K-Means clustering resulted in two small clusters and two larger clusters:

Cluster 1 – 3 sites, 84 species of fish – was dominated by species commonly associated with coral reefs

Cluster 2 – 8 sites, 173 species of fish – was dominated by species commonly found on trawl grounds

Cluster 3 – 20 sites, 225 species of fish – included some pelagic species as well as those found on trawl grounds

Cluster 4 – 23 sites, 271 species of fish – common trawl grounds species, but the dominant species were mostly different to those of cluster 3 and the pelagic species were not present.

Information on the most abundant species of fishes characterising these clusters is provided in Table 2.7-17.

The results of the LDA indicated that only 7 of the physical co-variates appeared to be important in influencing the distribution of the fish assemblage clusters, and were included in the discriminant functions (Table 2.7-18). The three co-variates that had the highest partial r^2 and F values, and therefore were most important in the discriminant functions were standard deviation of salinity (CARS_S_SD), the standard deviation of temperature (CARS_T_SD) and the amount of sand (auDB_SAND). The overall jack-knifed discrimination accuracy was ~79% correct and ranged from 75% correct for cluster 3, to 100% correct for cluster 1, which had only 3sites (Table 2.7-19).

Table 2.7-18. LDA statistics of co-variates included in the discriminant functions of the fish bycatch data.

Co-Variate	Partial r^2	F-value	Pr > F
CARS_S_SD	0.6959	38.13	<.0001
CARS_T_SD	0.4445	13.07	<.0001
auBD_SAND	0.3103	7.20	0.0004
CARS_PO4_AV	0.2632	5.24	0.0035
SW_CHLA_Y_SD	0.1731	3.28	0.0290
SW_D_B_IRRAD	0.1669	3.07	0.0369
CARS_O2_AV	0.1192	2.03	0.1231

As was done with the seabed habitat survey data, the LDA functions were used to predict the cluster membership of each 0.01 cell in the entire Torres Strait study area, based on the values of the co-variates at each cell. These predictions were then mapped to show the estimated distribution of benthic fish assemblages (Figure 9). The prediction beyond the trawl fish sampling area has unknown certainty and is likely to be unreliable.

Cluster 1 which was predominantly reef associated fish is within the area closed to the fishery. Outside the area sampled, the discriminant functions have allocated virtually the whole of the western Torres Strait to this cluster; something which is plainly incorrect. Interestingly though, the areas just to the

east of the outer Barrier reef have also been allocated to cluster 1 (Figure 2.7-14). Cluster 2 (predominantly trawl ground fish) is concentrated in the north east of the study area and to the eastern Torres Strait outside the study area. Clusters 3 & 4 extend north-south in the central part of the Strait.

Table 2.7-19. Cross-validation summary for the performance of the discriminant functions in assigning bycatch sites to clusters. The number in the top of each cell is the number of sites; the number in the bottom is the percent of the total number of observations for that cluster. The shaded diagonal cells show the numbers of sites that were correctly assigned to each cluster (columns, from rows).

CLUSTER	1	2	3	4	Total
1	3 100.0	0 0.0	0 0.0	0 0.0	**3** **100.0**
2	0 0.0	7 87.5	0 0.0	1 12.5	**8** **100.0**
3	1 5.0	1 5.0	15 75.0	3 15.0	**20** **100.0**
4	0 0.0	1 4.4	4 17.4	18 78.3	**23** **100.0**
Total	**4** **7.4**	**9** **16.7**	**19** **35.2**	**22** **0.7**	**54** **100.0**

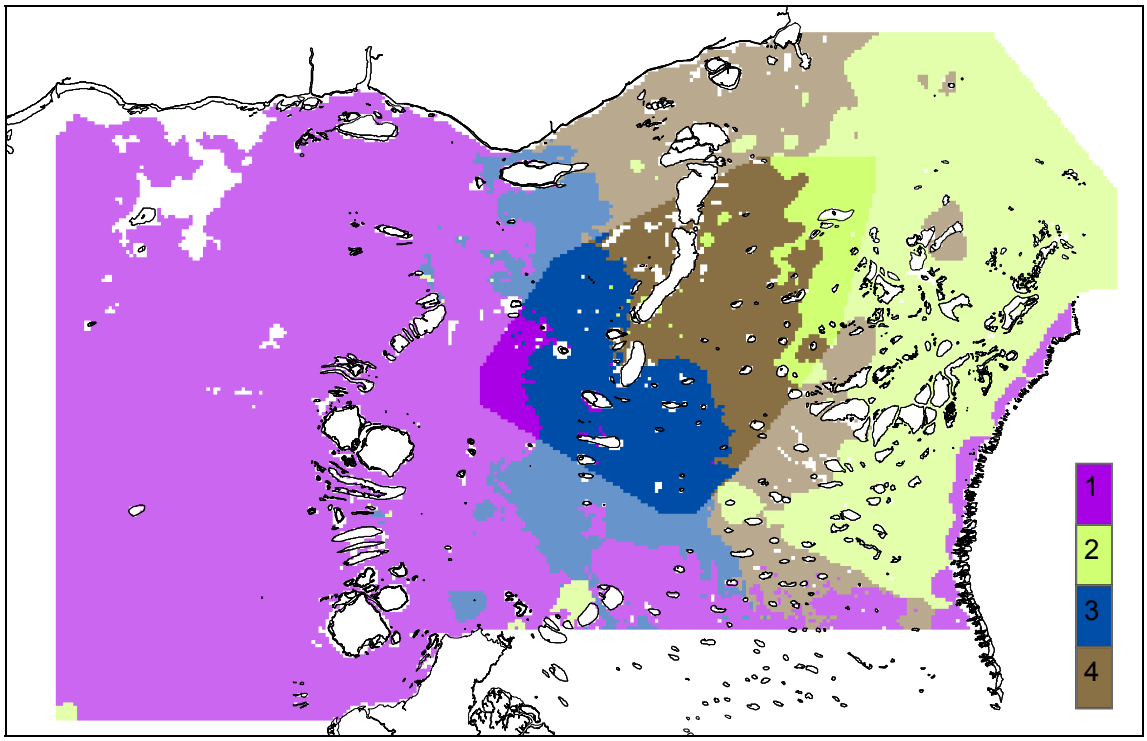


Figure 2.7-14. Mapping of predicted cluster membership (K-Means; 4 clusters) of the fish bycatch data for the 0.01 degree gridded physical environment data, using linear discriminant function analysis. The darker shaded mapping represents the predictions within the area where bycatch data was collected; the lighter shaded mapping is outside the survey area.

2.8. STRATIFICATION AND SAMPLING DESIGN

The future sampling for seabed biodiversity mapping in Torres Strait, as part of the CRC-TS, requires an optimal strategy for the selection of survey sites. The primary purpose of the survey itself is to obtain data on the spatial distribution of benthic biota, so that subsequent bio-physical modelling can make use of the physical environment co-variates to interpolate and map. Given that the number of sites that can be sampled is limited, it is obviously important to place the samples in a way that yields as much information as possible overall. This requires that the environment space, or multi-dimensional covariate space, rather than the 2-dimensional space must be sampled representatively and the approach to achieve this is stratification. Further, the stratification must be relevant to the benthic biotic, so it must be informed by measures of the biological importance of each covariate. This approach will optimally ensure that the biodiversity and physical attributes of as many different habitats types as possible, given the available resources, would be characterised. The physical variables collated as part of this project, which are known in advance of the survey, will be used to guide the stratification. Biological information will be taken into account by weighting the physical variables based on their relative importance in bio-physical relationships — variables of greater influence on biological patterns having a larger weighting and influence in the stratification.

From an earlier study (Pitcher et al, 2002) we have measures of the "importance" of these co-variates with respect to correlations with the abundance of many benthic species in a detailed survey of an adjacent area of the far northern Great Barrier Reef. Conceptually, important variables are those for which benthic composition changes significantly along a gradient of the variable. The survey should be designed to ensure that such important variables are sampled finely, so that the expected benthic diversity is reliably captured. That is, we should stratify our design with respect to the important variables.

Further, the sampling strategy should also consider the spatial resolution required for management utility. A scale of several 10s km was considered appropriate for broad scale characterisation. The implications of the spatial auto-correlation distance (section 2.6) and considerations of the benefit-cost of logistics (at about 1 site per hour) also indicate a sampling density of approximately 10 km average separation. In this approach, approximately 20-50 primary strata with similar physical characteristics will be identified from importance weighted physical covariates of almost fifty thousand 0.01° grid cells covering the shelf area of Torres Strait. The size (area) of strata will vary depending on the number of grid cells having particular similar physical characteristics. 'Replicate' future sampling sites, about 10-20, will be assigned to each primary stratum.

2.8.1. Stratification

The potential survey area in Torres Strait, after excluding reefs and other areas that were too shallow, included 41,285 cells of side 0.01° (~1.11 km), each square being a candidate sample site. Given the spatial autocorrelation distances, the average distance between sites should not exceed about 0.1° (~11.1 km) indicating that not less than about 400 of these squares should be sampled. A 10% margin was added to this lower limit, thus the design provided for 440 sites, although the resources of the future mapping project would allow only about two thirds of these to be sampled during 2003/04 to 2005/06. At the centre of each 0.01 degree cell, the values of 28 physical variables were collated or interpolated and represent the Torres Strait region as a cloud of 41,285 points within a 28-dimensional physical-variable space. Ultimately, this space was to be partitioned into 440 relatively homogeneous

regions (or *strata*), such that the expected benthic biodiversity would be homogeneous within each stratum but heterogeneous among strata. A sampling site would then be selected from each stratum to produce a set of 440 sites. This section describes the methods for achieving this partitioning or stratification of physical-variable space.

2.8.1.1. Principles of Partitioning

The basic principle behind the partitioning can be illustrated with the following simple two-dimensional example. Consider two physical variables x and y for which we have values at 1000 sites, and suppose that these sites sample the covariate space roughly uniformly (Figure 2.8-1(a)). We wish to partition the covariate space into 20 strata. If the two variables were equally important, then the partitioning in Figure 2.8-1(b) would be adequate, since the strata are roughly the same width in x and y. This partitioning was achieved using the "partitioning around medoids" (PAM) algorithm (Kaufman and Rousseeuw, 1990) (see below).

However, suppose the x variable is known to be 4 times more important than the y variable. Then we would prefer a partitioning more like that in Figure 2.8-1(c), where the strata are roughly 4 times narrower in the x direction than in the y. This is very simply achieved by first scaling the x variable 4-fold and then applying PAM to the scaled covariates, as in Figure 2.8-1(d).

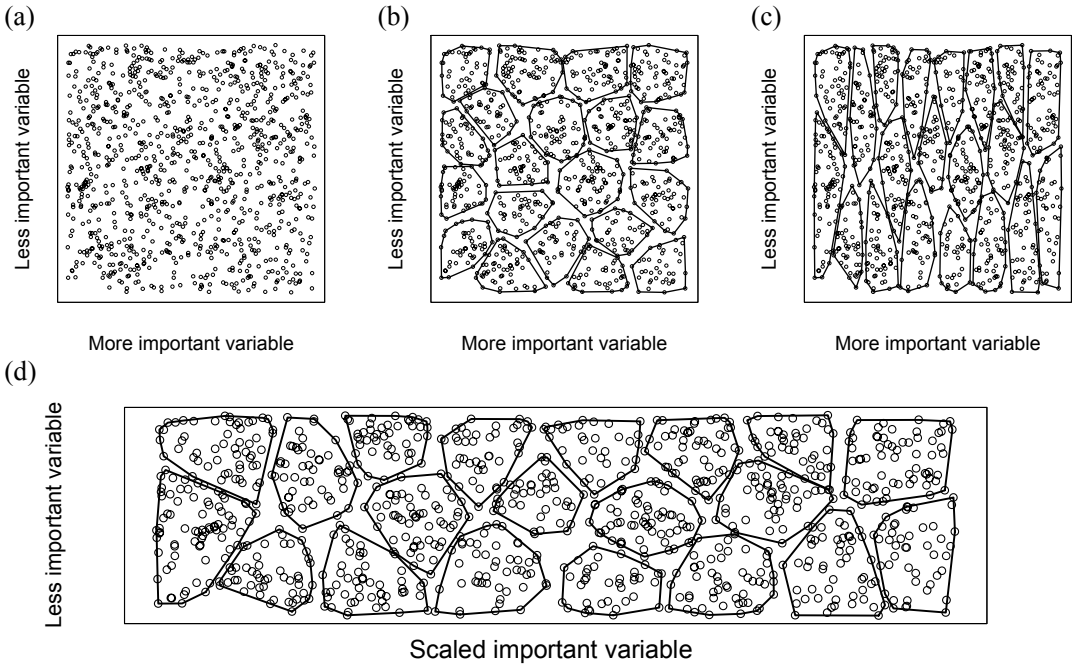


Figure 2.8-1. Partitioning covariate space in two dimensions: (a) 1000 points randomly sampled from the square covariate space. (b) a partitioning into 20 clusters using PAM; (c) a preferred partitioning that accounts for the relative importance of the variables; (d) the partitioning in (c) is achieved using PAM on the scaled covariate space.

The partitioning of the Torres Strait grid cells was an analogous procedure in 28 dimensions. Each variable was scaled so that its 'range' was proportional to its importance. However, unlike in the

example, the physical variables were not uniformly distributed across their range and may have extreme outlying values. To guard against the distorting influence of such values, the 'range' was taken as that of the middle 95 percentiles. The term "*I95R*" is used here for this range, in acknowledgment of the inter-quartile range, *IQR*, of which this is a generalization. Formally,

$$I95R(v) = v_{(97.5\%)} - v_{(2.5\%)},$$

where $v_{(i\,\%)}$ is the i-th percentile of variable v.

2.8.1.2. *Variable Importance*

The collated physical variables were quantified on various disparate measurement scales that were unlikely to have any direct relevance to their biological importance. To scale the variables appropriately to inform the stratification, it was necessary to derive an importance value for each variable. The primary component was the biotic importance, but it was also necessary to include a study area adjustment and a reliability adjustment. The biotic importance quantifies the link between the biota and the physical variables and was developed from the detailed species data sampled in the adjacent GBR, but was checked for consistency against the analyses conducted in section 2.7. The study area adjustment was a refinement to the biotic importance to account for potential differences in the range of the physical variables between the Torres Strait and the GBR. The reliability adjustment was a further refinement to reduce the influence of variables that are spatially poorly resolved. These are described in detail below.

Biotic importance I_{bio}

In a previous study, Pitcher et al (2002) performed univariate analyses of 30 benthic statistical assemblages (comprising ~800 species) and 90 single species analyses on 306 sites using a similar suite of physical covariates as explanatory variables. They derived tree models for abundance, logistic regression models for presence/absence data and lognormal regression models for abundance conditional on presence. Their method used model selection to arrive at parsimonious models with some explanatory power and lead to the derivation of a measure of importance for each variable. For each species the relative amount of variation explained by each variable was computed, i.e. the contribution of the variable to the overall R^2. The average of this quantity over all species was defined to be the importance for that variable.

Clearly, the actual dependence of biota on the physical variables is multivariate and highly complex. Moreover, the explanatory power of the physical variables is fairly low, averaging about 30%. Nevertheless, this definition of importance captured the broad pattern over a fairly diverse range of biota. Also it allowed for variation in explanatory power, since species that had low R^2 contributed less to the importance.

The three types of models considered by Pitcher et al were in broad agreement over the ranking of the variables. However, as the tree model approach was most readily cross-validated, these results are reproduced here; the importances are shown in Figure 2.8-2(a).

An alternative but similar approach called random forests (Breiman, 2001) was also considered. In this procedure a bootstrap sample (with replacement) of all 306 sites is taken and a full tree model is

fit without pruning. The method for selecting the splitting variable at each node differs from standard trees, where all variables are considered for splitting. In contrast, for random forests, a reduced set of *m* candidate variables, chosen at random, are considered for splitting, and the candidate with the best split is selected as usual. This bootstrap procedure is repeated 500 times to produce a 'forest' of tree models. Predictions can be made from the forest by taking the average prediction from the individual trees. For each sample, roughly 37% of sites are not selected. These 'out-of-bag' sites provide a test data set for estimating (without bias) the prediction error of the forest as a whole. As *m* increases two effects occur: the prediction error of individual trees improves, and the correlation among trees increases. The first acts to reduce overall prediction error but the second acts to increase it. There is therefore an optimal value for *m*, which Breiman has shown to be close to the square root of the total number of variables. Given that 28 co-variates were available for Torres Strait, $m = 5$ was chosen.

The out-of-bag sites also provide a means of defining importance. The importance of variable *v* is the percent rise in the out-of-bag mean sum-of-squared errors when the values of *v* are randomly permuted. This is a relative measure that can be averaged over species. The results are shown in Figure 2.8-2(b).

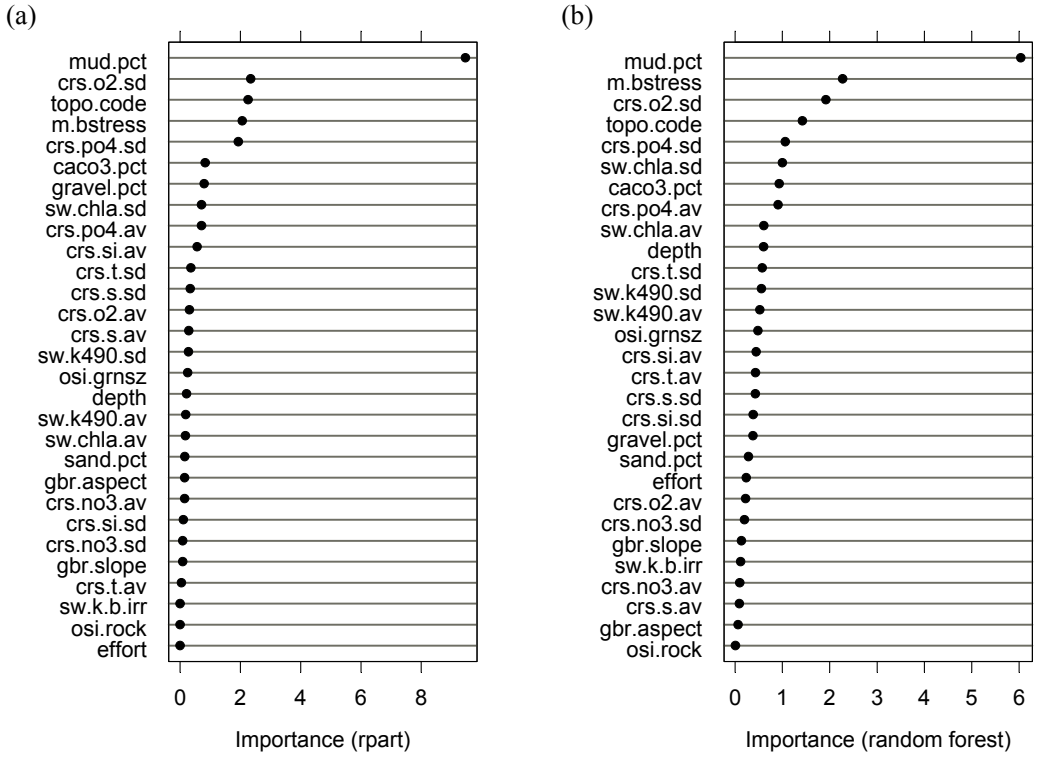


Figure 2.8-2. Variable importance computed by (a) cross-validated trees and (b) random forests.

The results for random forests were qualitatively similar to those for the tree models with slight adjustments to the rankings. The decay in importance with ranking was somewhat smoother for the random forests. Also, because of the use of random candidate variables, the random forests procedure tended to overcome the potential of some variables to dominate other closely correlated variables in

the fitting; each variable gets a 'fair go'. Thus, the random forest importances were considered more robust and were used in the stratification approach.

Two variables in the Torres Strait dataset, slope and topographic code, were considered too unreliable and were excluded from the stratification. A third variable, % rock from the OSI auSeabed database that was considered unreliable and found to have very low importance in the in the northern GBR, nevertheless proved moderately important in the Torres Strait from the limited biotic information available (section 2.7). Over several analyses, on average it ranked about the same importance as other OSI sediment attributes, and, as we did not have an importance measure for this variable, we arbitrarily assigned the same importance to OSI rock.

Table 2.8-1. Calculation of adjusted importance I_{adj}: d_{err} is error distance in degrees, I_{bio} is the random forests biotic importance, reliability is $R = (d_{err})^{-½}$, and $I_{adj} = (I_{bio}QR)^{0.6}$.

Variable	Distance d_{err} (°)	Biotic imp. I_{bio}	I95R ratio Q	Reliability R	Adjusted imp. I_{adj}
m.bstress	0.008	2.3	2.6	11.5	12.6
sw.k490.av	0.004	0.5	3.2	16.1	7.3
osi.mud	0.045	6.0	0.9	4.7	7.2
sw.k490.sd	0.004	0.6	2.1	16.1	5.8
sw.chla.sd	0.004	1.0	1.0	16.1	5.3
sw.chla.av	0.004	0.6	0.7	16.1	3.1
agso.dem2	0.037	0.6	1.9	5.2	2.9
cars.o2.sd	0.352	1.9	1.5	1.7	2.6
cars.po4.av	0.912	0.9	3.5	1.0	2.1
osi.crbnt	0.110	0.9	1.0	3.0	1.8
osi.grnsz	0.045	0.5	1.1	4.7	1.7
osi.gravel	0.045	0.4	1.4	4.7	1.7
cars.s.sd	0.312	0.4	3.0	1.8	1.7
cars.po4.sd	0.912	1.1	2.0	1.0	1.6
cars.s.av	0.312	0.1	11.6	1.8	1.5
effort	0.039	0.2	1.4	5.1	1.4
osi.sand	0.045	0.3	1.1	4.7	1.2
sw.d.ben.irr	0.020	0.1	1.4	7.0	1.1
cars.si.av	0.362	0.4	1.5	1.7	1.1
cars.t.sd	0.312	0.6	1.1	1.8	1.1
cars.si.sd	0.362	0.4	1.1	1.7	0.8
dem.aspect	0.037	0.1	1.3	5.2	0.6
cars.no3.sd	0.347	0.2	1.3	1.7	0.6
cars.no3.av	0.347	0.1	1.0	1.7	0.3
cars.o2.av	0.352	0.2	0.4	1.7	0.3
cars.t.av	0.312	0.4	0.1	1.8	0.3
dem.slope	0.037	0.1	0.0	5.2	0.0

Study area adjustment Q

The raw importance values from the GBR needed to be adjusted to take into account that the Torres Strait study area is different. Some variables, in particular bottom stress, have a larger range over Torres Strait than over the far northern GBR survey area. Such variables may therefore be more important in Torres Strait. Thus, importances were rescaled in proportion to the ratio of *I95R* between the two regions; the scale factor Q (see Table 2.8-1).

The derived importances were also checked by comparisons with analyses of biotic data from the Torres Strait (section 2.7). It was not possible to perform an importance analysis for Torres Strait in the same detail as for the northern GBR study, because the Torres Strait datasets largely consisted of very generalized habitat characterisation, or for the single species-biomass dataset (trawl fishes) very limited coverage. However, a guide to relative covariate importance was available from *F*-values from stepwise discriminant analysis on clusters defined by substrate type or habitat type (section 2.7). The selected variables were in broad agreement with the adjusted importances here. In particular, bottom stress was identified as important for differentiating both substrate and habitat.

Reliability adjustment R

The third consideration was that the physical variables had widely differing reliability that needed to be taken into account in the calculation of importance. All the physical variables were available on the design grid of 0.01° cells. However, most variables were interpolated onto this grid based on sample data at a coarser resolution. Therefore, an error distance d_{err} was defined to quantify this spatial imprecision (see Table 2.8-1).

The CARS data were interpolated from a rather limited number of CTD casts, and, for example, in the case of silicate, the average density of casts with silicate data was approximately 1 in 1,300 km², corresponding to an average distance d_{err} of 0.36 degrees between casts. For the effort data, which came from logbooks reporting effort at 6-minute resolution, d_{err} was set to be the average distance from the design grid cell to the centre of the 6-minute effort cell. For the OSI data, d_{err} was set to be the average distance to a sample point from each design grid cell. The SeaWiFS data in their raw form were already specified at the same scale as the design grid; in this case d_{err} was set to be the average distance to the grid cell centre within a grid cell.

The ratio of largest to smallest d_{err} was about 230 (refer Table 2.8-1). It was considered that rescaling over such a large range would be too severe an adjustment and would effectively eliminate the CARS variables from influencing the stratification. Thus, the square root of d_{err} was taken and its reciprocal was defined as the reliability scaling factor R.

Adjusted biotic importance I_{adj}

To incorporate reliability, initially the product $I_{bio}QR$ was considered and compared with the study area-adjusted importance $I_{bio}Q$. First, the two adjusted importances were normalized to sum to 1 and sorted in descending importance, as in Figure 2.8-3. The reliability-adjusted importance has much stronger contrast between low-ranked and high-ranked variables, a distortion which was considered unacceptable. Therefore, the reliability-adjusted importance was ‘tuned’ by raising to a power γ. The value of γ was chosen to make the tuned importance match the study area-adjusted importance as

closely as possible: $\gamma = 0.6$ gave the minimum sum-of-square differences (compare the blue and green lines in Figure 3):

$$I_{adj} = (I_{bio}QR)^{0.6}.$$

Finally, for each physical variable v, the scaled version v_{scaled} that was used in the stratification was defined thus:

$$v_{scaled} = [v \div I95R(v)] \times I_{adj}(v)$$

This scaling ensures the *I95R*'s of the scaled variables are proportional to the adjusted importances.

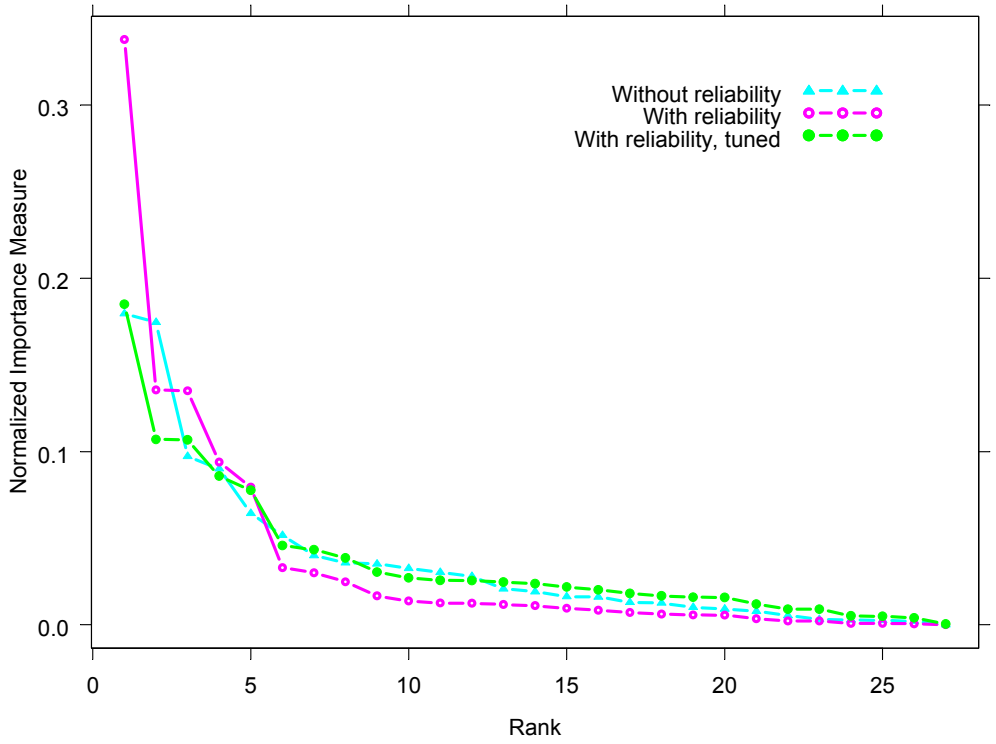


Figure 2.8-3. Importance measures excluding reliability ($I_{bio}Q$), including reliability ($I_{bio}QR$), and including reliability, but tuned to match the shape without reliability $(I_{bio}QR)^{0.6}$. Each version is normalized to sum to 1. The orders of the variables with and without reliability are different.

2.8.1.3. The Clustering Process

Having achieved a biologically informed scaling of the physical variables, the next step was partitioning. However, before proceeding, it was necessary to reduce the dataset for computational manageability and to provide an orthogonal coordinate space for clustering.

There was a certain degree of redundancy among the physical variables. For instance, some variables (phosphate, silicate, chlorophyll A, K490) had a high correlation (>80%) between their average value and standard deviation. There was strong correlation (>77%) among all SeaWiFS chlorophyll A and

K490 measurements, and there were also some negative correlations, e.g. between temperature and silicate standard deviations (–85%). Hence, there was an opportunity to apply data reduction techniques to make the data set more manageable and, importantly, orthogonal prior to clustering.

Table 2.8-2. Variable loadings for the first 7 principal components. Absolute loadings greater than 0.5 are highlighted in yellow, and absolute loadings between 0.3 and 0.5 are highlighted in green. The variables are ordered by adjusted importance. Relative variance is the fraction of the total variance explained by the principal component.

Loadings	Principal Component						
Variable	**1**	**2**	**3**	**4**	**5**	**6**	**7**
m.bstress	**+0.80**	+0.41	–0.41	+0.02	+0.10	–0.04	–0.04
sw.k490.av	+0.31	–0.47	+0.29	–0.12	**+0.61**	–0.12	+0.19
osi.mud	–0.14	–0.49	**–0.79**	–0.28	–0.02	+0.01	+0.07
sw.k490.sd	+0.27	–0.41	+0.13	+0.19	+0.02	+0.30	**–0.52**
sw.chla.sd	+0.32	–0.36	+0.09	+0.28	**–0.67**	+0.08	+0.16
sw.chla.av	+0.16	–0.19	+0.09	–0.08	–0.19	–0.37	+0.34
agso.dem	–0.08	+0.04	–0.12	+0.35	+0.23	+0.44	+0.36
cars.o2.sd	+0.11	+0.00	+0.12	–0.38	–0.05	+0.49	+0.27
cars.po4.av	+0.05	+0.09	+0.12	–0.44	–0.10	+0.07	–0.06
osi.crbnt	+0.03	+0.06	+0.06	–0.16	–0.07	–0.26	+0.04
osi.rock	+0.04	–0.01	+0.01	+0.04	–0.03	+0.00	+0.04
osi.grnsz	–0.02	–0.06	–0.06	–0.12	+0.03	+0.08	–0.20
osi.gravel	+0.02	+0.06	+0.04	+0.12	–0.02	–0.12	+0.38
cars.s.sd	+0.07	–0.03	+0.08	–0.28	–0.08	–0.07	–0.01
cars.po4.sd	+0.03	+0.06	+0.08	–0.29	–0.14	+0.07	–0.07
cars.s.av	–0.05	+0.08	–0.05	+0.11	–0.13	+0.17	+0.00
effort	–0.03	–0.01	–0.05	+0.00	–0.02	–0.08	–0.14
osi.sand	+0.00	+0.01	+0.06	–0.05	+0.01	+0.07	–0.27
sw.d.ben.irr	+0.01	+0.01	+0.03	–0.15	–0.09	–0.08	–0.11
cars.si.av	–0.02	–0.05	–0.03	+0.10	+0.03	–0.21	–0.13
cars.t.sd	+0.03	+0.05	+0.06	–0.20	–0.09	+0.14	+0.05
cars.si.sd	–0.02	–0.04	–0.03	+0.13	+0.03	–0.23	–0.08
dem.aspect	+0.00	+0.00	+0.00	–0.01	+0.00	–0.01	+0.00
cars.no3.sd	+0.01	+0.01	+0.01	–0.10	+0.01	+0.21	+0.11
cars.no3.av	+0.01	–0.01	–0.01	–0.01	+0.01	+0.05	+0.04
cars.o2.av	+0.01	–0.01	+0.01	–0.01	+0.01	+0.01	+0.04
cars.t.av	+0.01	+0.00	+0.01	–0.03	–0.02	+0.00	–0.02
Relative Variance	**0.52**	**0.21**	**0.12**	**0.04**	**0.03**	**0.01**	**0.01**

Data reduction

Singular value decomposition (SVD) was used to separate the data into principal components, from which we retained the most important components accounting for 99% of the variance in the data. This was contained in the first 14 components, and in fact the first 7 components contained 95% of the

variance. SVD decomposed the 41,285 × 28 data matrix X of scaled physical variables into a product of matrices UDV^T, where U was the 41,285 × 28 score matrix, D was the 28 × 28 diagonal matrix of singular values, and V was the 28 × 28 orthogonal loadings matrix. To project the data into a smaller dimensional space, but retain the relative distances of the data, a new data set was defined as UD^* where D^* (28 × 18) consists of the first 18 columns of D. This data is equivalent to rotating the scaled data by V (i.e. XV) and projecting into the 18-dimensional subspace spanned by the first 18 columns.

The effect of this transformation was observed by examining the variable loadings V. The rows of V correspond to the original variables and the columns to the principal components. Large values (on the scale 0 to 1) indicate alignment of the variable with the principal component. The important variables should be expected to have high loadings on the first few principal components, and the less important variables to have higher loadings on the later principal components.

The loadings on the first seven principal components are shown in Table 2.8-2. Principal component 1 was mainly associated with bottom stress, whereas the second component was associated with various SeaWiFS measurements, as well as with bottom stress and mud. Because the 3 most important SeaWiFS variables are highly correlated with one another, they have similar loadings. The 3rd component was principally mud, the 4th introduced depth and two of the CARS variables, and the 5th component comprised mainly the difference between average chlorophyll A and K490.

Including geographic constraints

Another important consideration was whether spatial position should be included in the stratification. In the absence of covariate information, it would be usual to stratify entirely on geographical position, making each stratum simply connected. On the other hand, if we ignore geography completely, and base the stratification only on physical covariates, then the strata will tend to be fragmented in geographical space. This would not necessarily be a bad thing. However, if the fragments become very small then the quality of the stratification may become degraded by spatial uncertainty in the covariates themselves.

This issue was examined in the design of the GBR Seabed Mapping survey and the conclusion was that it would be prudent to include a small amount of geography (Pitcher et al, 2002). Using the recommendations from that study, latitude and longitude were scaled equally so that the *I95R* of the scaled latitude equalled 0.25 times the *I95R* of the first principal component of the rotated data. The scaled spatial variables were included as extra dimensions in the clustering, and their effect was generally to prevent the clusters becoming too highly fragmented in space.

The PAM and CLARA algorithms

The clustering algorithm “partitioning around medoids” (PAM) of Kaufman and Rousseeuw (1990), which is implemented in Splus, was used to cluster the physical dataset. The PAM algorithm is a robust alternative to the k-means algorithm. It uses a distance matrix and the number of clusters must be specified. Whereas K-means minimizes distances to the average for the cluster, in PAM, each cluster contains a *medoid* that is the cluster member whose summed distance to all other cluster members is a minimum. The medoid is a kind of generalized median for multiple dimensions; it is to this that the algorithm owes its robustness. The algorithm works by searching for clusters that minimize the total distance to cluster medoids.

PAM is not immediately useable for large data sets, because the size of the distance matrix becomes unmanageable. Therefore Kaufman and Rousseeuw's CLARA algorithm, which is an implementation of PAM for large data sets, was applied. This works by first selecting a random subset of the data, then applying PAM to generate a clustering, and finally assigning the remainder of the data to the nearest cluster in the subset. The procedure is repeated many times to give several candidate clusterings, from which the candidate that minimizes the total distance to cluster medoids is chosen. The algorithm can be tuned by adjusting the subset size and the number of repeats, both of which should be as large as practicable.

Further, a weighted version of CLARA was developed specifically for this project. In this implementation, each initial subset was selected with non-uniform probabilities or weights, which enabled the clustering to be influenced to some extent to seek rarer physical environment strata, as explained below.

Two-stage partitioning

The partitioning was performed in two stages. In stage 1, we generated an initial coarse partitioning of the entire data set into 50 'superclusters', or primary strata. Then in stage 2, each supercluster in turn was partitioned, generating a total of 440 subclusters.

The initial reason for having two stages was computational efficiency. For k clusters and n observations, the computation time is of order kn^2; but if $\sqrt{k}$ superclusters was computed first, and then $\sqrt{k}$ subclusters (on average), the computation time can be reduced to the order $\sqrt{k}\, n^2$. In fact stage 1 is the most computationally intensive stage, taking of order $\sqrt{k}$ times longer than stage 2. Even for 50 superclusters, which was somewhat larger than $\sqrt{440}$, the computational saving was substantial. This was an important consideration when developing a method, particularly where many subsets of the data must be run.

However, the main reason for using a two-stage method was that it allowed more control over the partitioning. This was because at stage 2, it becomes possible to choose the number of subclusters within each supercluster, subject to a total of 440. In particular, it was possible to raise the level sampling effort into uncommon and rarer areas in covariate space, that may be potentially more interesting in terms of biota, at some expense to common areas.

Choosing the number of subclusters

After stage 1, there were 50 superclusters of various sizes ranging from 62 to 2113 cells. Then it was important to determine how to optimally distribute the 440 subclusters among the 50 superclusters.

In order to answer this question, initially the following hypothesis was adopted: clusters with large numbers of cell members tend to be more homogeneous and represent commonness, compared with small clusters. Support for this hypothesis can be seen in Figure 2.8-4 for a synthetic bivariate normal data set. The larger clusters (in terms of numbers of cells) near the middle have smaller bivariate space (i.e. are more homogeneous), whereas the more heterogenous clusters around the fringes tend to have fewer points (i.e. are smaller clusters).

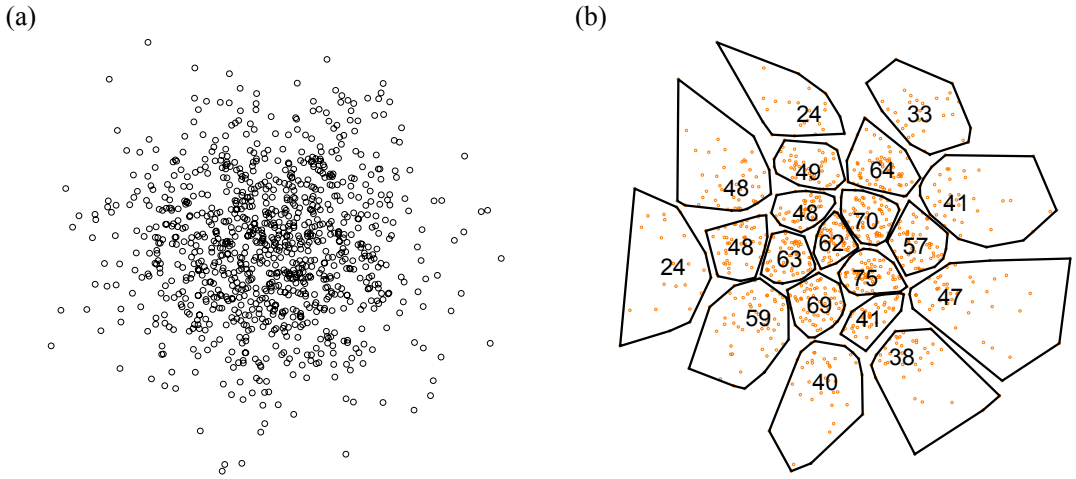


Figure 2.8-4. (a) Bivariate normal distribution of 1000 points. (b) Partitioning into 20 clusters using PAM. Each cluster is labeled by the number of points in the cluster. The more populous clusters tend to be tighter and so more homogeneous.

Therefore the stratification strategy should be such that the density of sampling should be lower for larger superclusters, i.e. the number of subclusters N_{sub} depends sub-linearly on the supercluster size S. This issue also arises in the context of species-area curves, where the number of species increases with area sampled, but less than linearly. In fact, for species-area curves a square-root relationship is sometimes used. Following this principle, the initial approach could be $N_{\text{sub}} \propto \sqrt{S}$.

This approach would attempt to bias the sampling away from common sites towards rarer, perhaps more 'interesting', sites so that they also can be sampled adequately. Nevertheless, the square-root approach provides a somewhat crude approximation to the amount of 'interest' in a supercluster, relating it simply to the size of the supercluster, without regard to its contents. A better approach would be to quantify the interest as a sum over the interest in individual sites. For this, it was necessary to define the interest at a site.

The more common sites are those lying in high-density areas of covariate space. Since common sites will be well sampled in any case, it was reasonable to define 'interest' as some inverse power of density. However, computing the density in more than 2 dimensions is difficult; instead the one-dimensional densities of each physical variable was considered separately. Suppose d_{vi} is the density of variable v at site i, normalized so that the total density over all sites is 1. Then we define the interest w_i at site i as the variable importance-weighted sum,

$$w_i = \sum_{v=1}^{28} I_{\text{adj}}(v) d_{vi}^{-a} ,$$

where $a > 0$ is a parameter to be chosen. Then define the interest of a supercluster as the total interest over sites within the supercluster, and choose the number of subclusters to be proportional to this quantity. That is, for the k^{th} supercluster $C(k)$:

$$N_{\text{sub}}(k) \propto \sum_{i \in C(k)} w_i .$$

The density is estimated from the 41,285 values using a gaussian kernel whose width is calculated by biased cross-validation (Scott, 1992). As an example, Figure 2.8-5 shows the true density (total area = 1) for bottom stress. The bulk of the distribution lies below 0.5; whereas previous experience has demonstrated that sites above 0.7 were of particular interest for epibenthic fauna (see section 2.7).

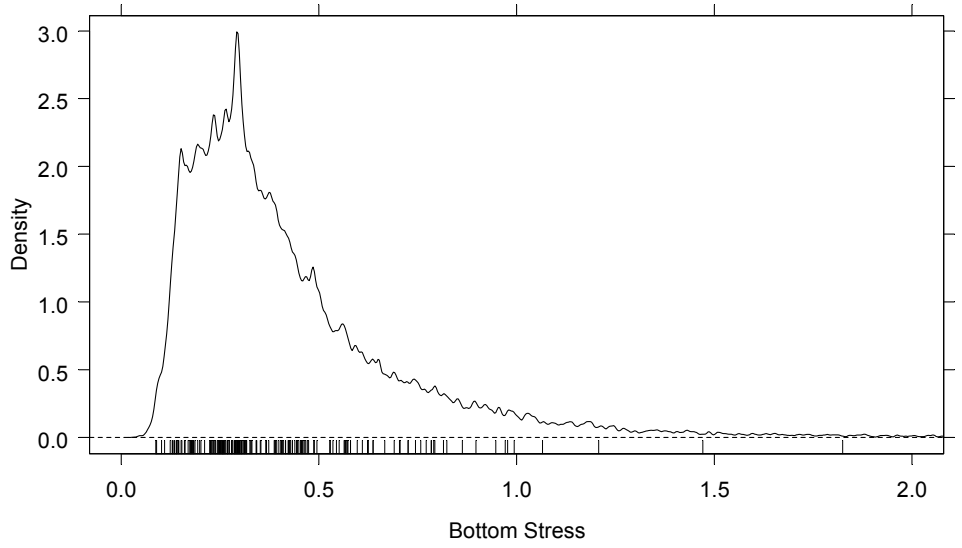


Figure 2.8-5. Density of bottom stress estimated by a gaussian kernel of width 0.018 calculated using biased cross-validation. Also shown is a 'rug' of values for 200 randomly selected sites.

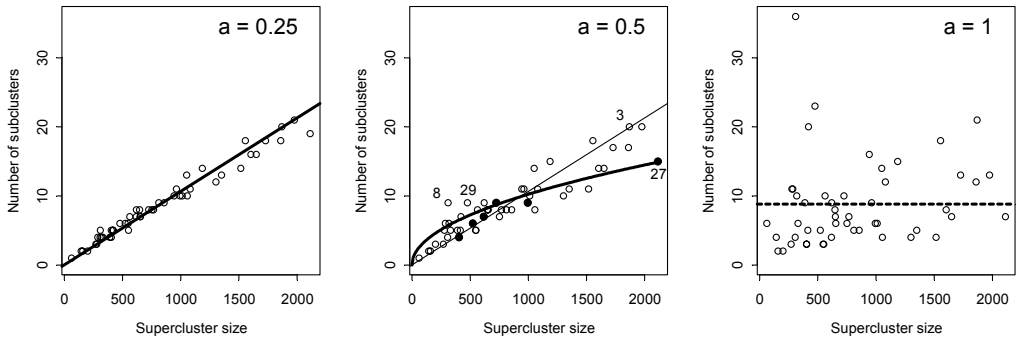


Figure 2.8-6. Number of subclusters vs supercluster size for 3 different values of the exponent *a*. The sloping line corresponds to $N_{sub} \propto S$, the curve to $N_{sub} \propto \sqrt{S}$, and the horizontal line to N_{sub} = constant. In the middle plot, four superclusters are labelled for later reference in the text, and superclusters denoted by a black dot are mapped in Figure 2.8-8.

Figure 2.8-6 shows the relationship between number of subclusters and supercluster size for a = (0.25, 0.5, 1). For the case a = 0.25, the relationship was almost linear; this was barely distinguishable from the case a = 0, in which all sites had equal interest. At the other extreme, case a = 1 flattened the relationship, making number of subclusters nearly independent of supercluster size and too sensitive to individual high-interest sites within a supercluster. The intermediate case a = 0.5 was close to the

square-root proposal discussed earlier and provided the required increased sampling of rarer sites without unacceptable under-sampling of common sites. This value for a was used as it provided an improved stratification adjustment compared with the initial square-root proposal.

There was a concern that, at the superclustering stage, rarer sites might be missed in the CLARA random subset selection stage since rare sites would be unlikely to be selected in a small random subset and, as a consequence, the superclusters could be too large and homogeneous. Such superclusters, being comprised largely of common sites, would have fewer subclusters, and so there would be less chance of isolating the rarer sites into their own subclusters. Two steps were taken to reduce this risk. Firstly, we computed more superclusters than was computationally optimal (ie. $50 > \sqrt{440}$). Thus, superclusters would be smaller, allowing for better detection of heterogeneity within a supercluster. Secondly, a weighted version of CLARA was developed, with site interest w_i as the weighting. Thus, rarer sites were more likely to have a chance at being chosen in the random sample of the algorithm, and therefore more likely to seed a separate supercluster.

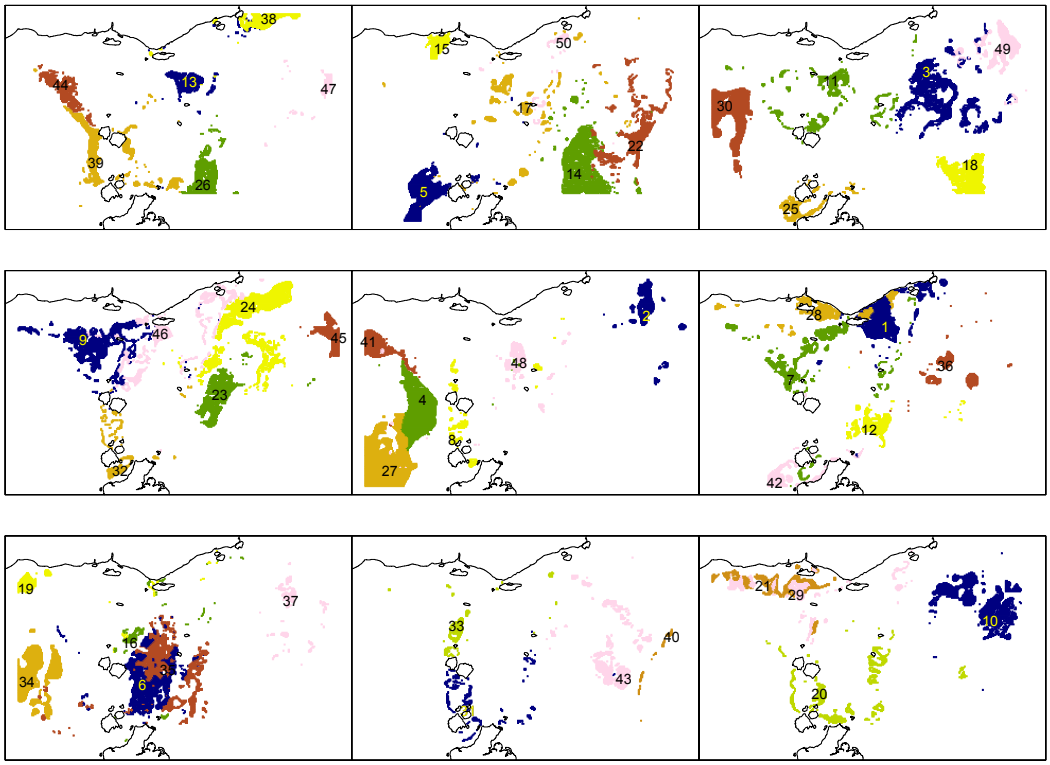


Figure 2.8-7. The 50 superclusters in geographical space. The clusters have been separated into nine panels in order to make them distinct. The largest cluster (27) is in the centre panel and the smallest (40) is in the bottom centre panel. Clusters 3 (top right) and 24 (centre left) have the largest number of subclusters (20).

Figure 2.8-7 shows maps of the resulting 50 superclusters after the first stage of clustering. Because the clustering was in covariate space, there was no guarantee that the clusters would be simply connected in geographical space, even though latitude and longitude were included as covariates. Indeed some clusters, especially those around the centre and north, were highly fragmented (e.g. 17,

11, 46 and 20). Despite their geographical appearance, these clusters' sites have similar physical characteristics. In the other hand, some clusters, especially those in the west, southwest and southeast, are fairly spatially contiguous (e.g. 14, 30, 18 and 27). Part of the reason for this is that the covariate values in these regions are based on spatial interpolation from sparse data points, and so the covariates vary smoothly in space.

Figure 2.8-8 and Figure 2.8-9 show the subclustering within some of the superclusters. In Figure 2.8-8, all the superclusters have an 'average' amount of subclusters, as indicated by the middling locations of the black dots within the vertical spread in Figure 2.8-6; the exception is supercluster 27, which has below-average sampling density. All these superclusters are spatially fairly contiguous, and the same is also true of their subclusters.

The cluster medoids (indicated by red dots) are the most central member of the cluster in covariate space. Often, but not always, the medoid is close to the geographical centre of the cluster. The medoid is a very useful by-product of the PAM/CLARA algorithm and, because it is always a cluster member, the medoid can be used as a representative of the cluster. This is a distinct advantage over the mean, which, for an irregularly shaped cluster, might not lie near any of its members.

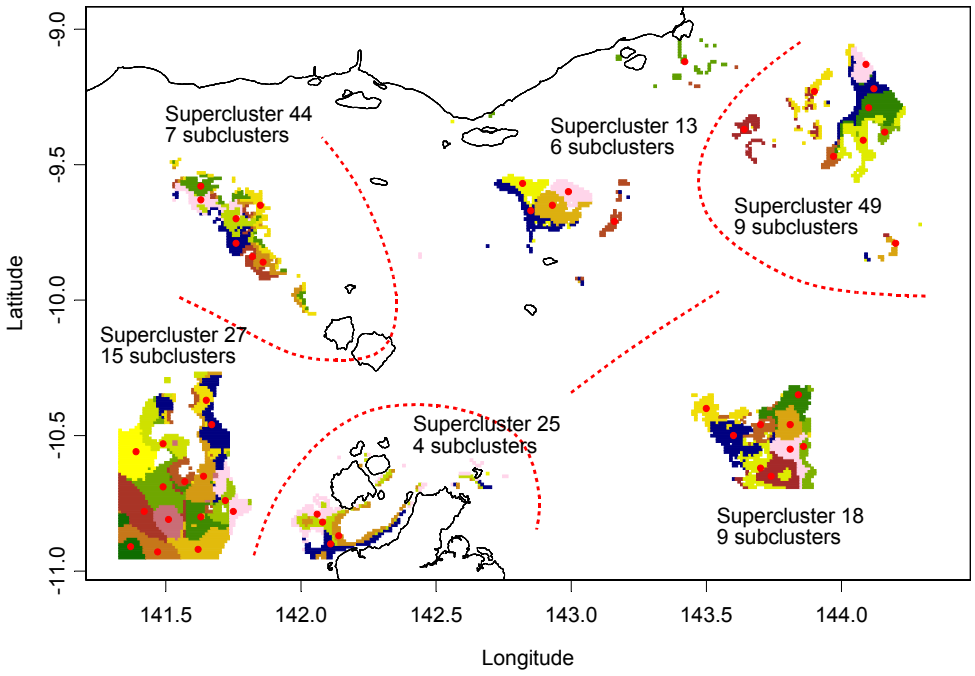


Figure 2.8-8. Six fairly compact superclusters in geographical space and the subclustering within them. The subcluster medoids are indicated by a red dot.

Three superclusters (3, 8 and 29) with above-average sampling (at top of the vertical spread in Figure 2.8-6) are mapped in Figure 2.8-9. These superclusters are much more fragmented, and some of the subclusters are also fragmented. Some of these clusters lie in the part of covariate space with high bottom stress, and, since bottom stress varies rapidly in geographical space, the clusters themselves

fragment over smaller scales. In particular, for supercluster 8, the clustering has four distinct well separated geographic areas and assigned 2 to 3 subclusters in each.

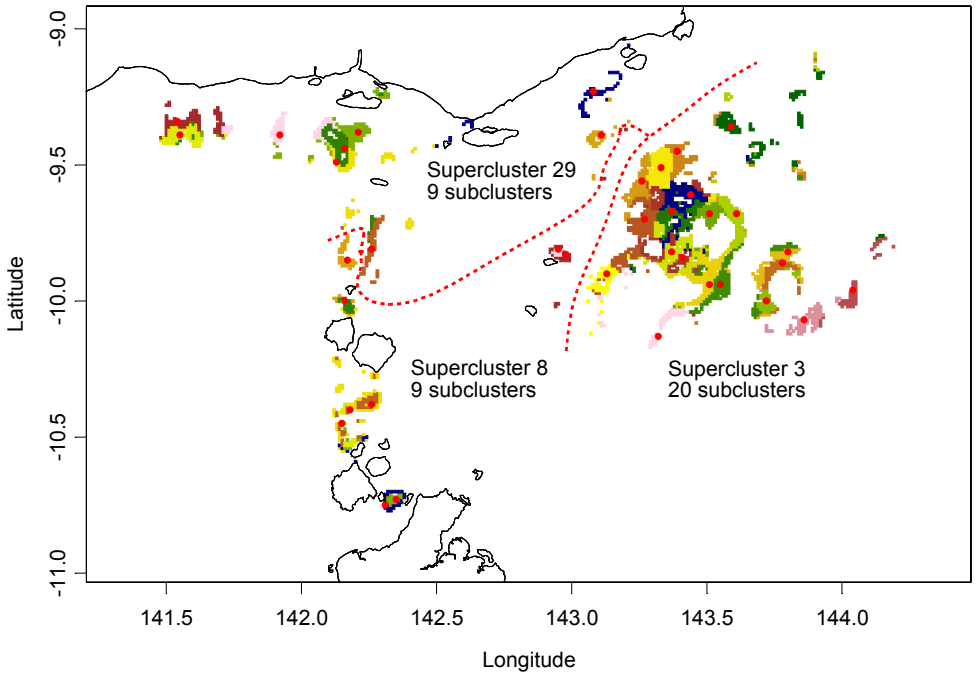


Figure 2.8-9. Three fairly fragmented superclusters in geographical space and the subclustering within them. The subcluster medoids are indicated by a red dot.

Assessing the resulting stratification

There is no unequivocally optimal approach to survey design. For instance, in the two-dimensional example of Figure 2.8-1, we could have used the k-means algorithm instead of PAM, and the resulting partitioning, which would have been different, would nevertheless have been a quite reasonable alternative. Although there is no single 'right answer', it is nevertheless necessary to establish that the resulting partitioning is reasonable. There are several ways to assess the stratification.

First, the strata were mapped. We have already partially shown this in Figure 2.8-7 to Figure 2.8-9. However, a map of all 440 strata would be rather overwhelming and very difficult to interpret. A clearer alternative was to plot the locations of the stratum medoids, since each medoid was in some sense the most typical representative of the stratum. In fact, the choice of medoids as actual survey sites would be a quite reasonable candidate sampling strategy and could be called "medoid sampling".

Figure 2.8-10 shows the medoid sites against the background of all possible sites. This would provide acceptable general coverage of the entire Torres Strait region. However, the sampling would be finer in some parts (such as the north, the east and around the longitude of Thursday Island) and coarser in other parts (the west, southwest and southeast). This was consistent with expectations and a desirable property of the stratification, which was being sought. The more heavily sampled regions are areas with either high bottom stress (around the islands and outer reef) or high chlorophyll A (the north), both highly important variables.

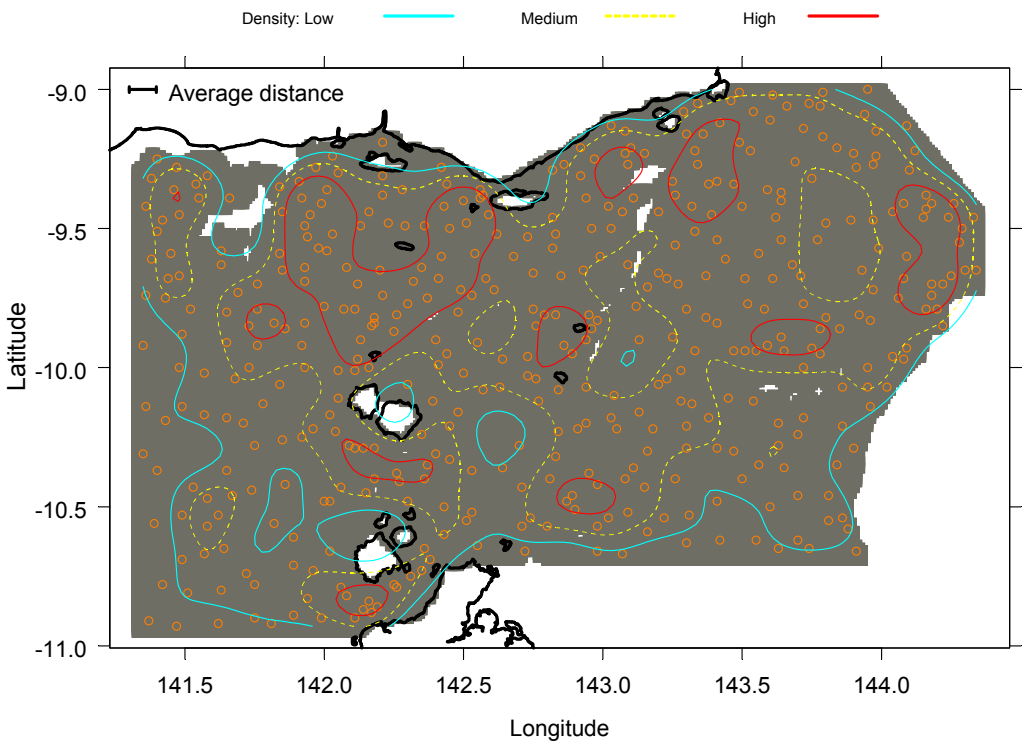


Figure 2.8-10. Medoids of the 440 strata. The bar labelled 'average distance' is the minimum spacing that would lie between 440 points if they were regularly spaced. The contour lines show the kernel density estimate of the medoids. The low, medium and high densities are, respectively, 0.66, 0.88 and 1.10 points per average distance squared. The 41,285 possible survey sites are indicated by the grey background.

The second way to assess the stratification was to examine the expected distribution of the physical covariates at the sample sites. Again, the medoid sampling can be used as a representative sampling. Figure 2.8-11 shows the density of bottom stress, average K490 and percentage mud over the stratum medoids compared to over all 41,285 sites. Transformed scales have been used, on which the distributions were roughly symmetrical, to make the comparison clearer. For completely random sampling, the density would be similar to that over the full data set. But in the medoid sampling, there was relatively less sampling in the high density (common) areas, and more sampling in the tails (rarer areas), which was the objective of the stratification. For bottom stress, more sampling is put into sites with values above 0.7, at the expense of the more common sites with values in the range 0.25–0.5. For K490, sacrifices are made in the range 0.075–0.1 to increase the sampling both above and below this range. For mud, the sampling is increased above around 40% at the expense of areas with no mud at all.

The representativeness of the medoid sampling can be checked by comparing its density with densities arising from many random samplings of the stratification. Figure 2.8-11 also shows confidence intervals for the density, which were obtained from the 5^{th} and 95^{th} percentiles of the pointwise densities of 20 random samples. Although there were small biases, overall the medoid-sampling density was fairly representative of the range of possible densities arising from stratified sampling.

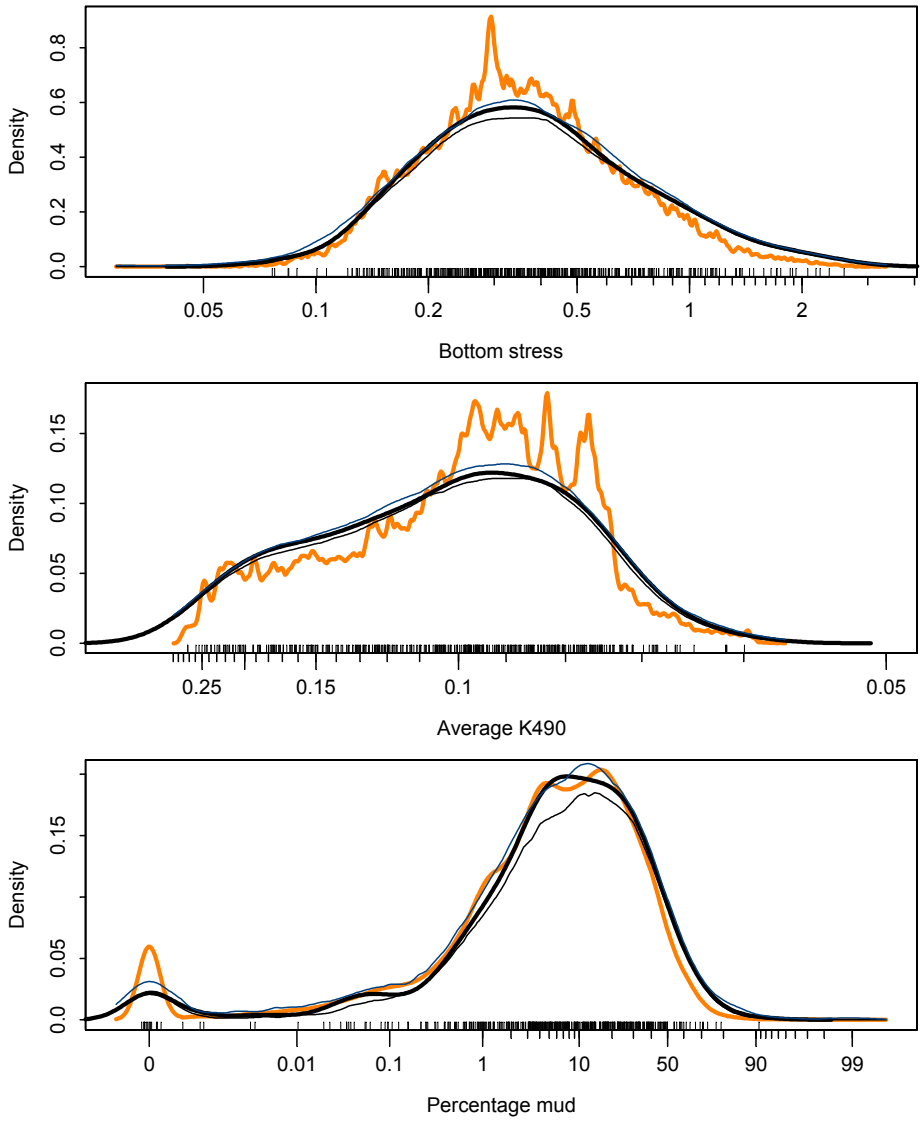


Figure 2.8-11. Distribution of the most important physical covariates on the subcluster medoids (black) and on the full Torres Strait data (orange). The thin curves are 90% confidence intervals for the density. For clarity we show covariates on a log scale for bottom stress, an inverse scale for K490 and a logit scale for mud. Also shown is a rug of the 440 medoid values (jittered around 0 for mud).

A similar approach was applied to the spatial distribution of medoid sites, by averaging the kernel density over repeated samplings (Figure 2.8-12). The results again show that the medoid sampling was fairly representative, although the medoid sampling was slightly denser around Thursday Island and just northeast of Cape York.

As a final assessment, the relationship between covariate value and the probability of selecting a site (the reciprocal of the stratum size) was examined. This is shown in Figure 2.8-13 for bottom stress, average K490 and percentage mud. Although there was considerable scatter in the probabilities, sites with rarer values in these covariates tended to have higher probability of selection. This was confirmed by the locally smooth regression (which has been applied to all 41,285 points, not just to those displayed). This also shows that the rarer physical environment combinations tended to reside in small

strata (<100 cells), whereas large strata (~250 sites) hold the more common sites. This is especially evident for bottom stress.

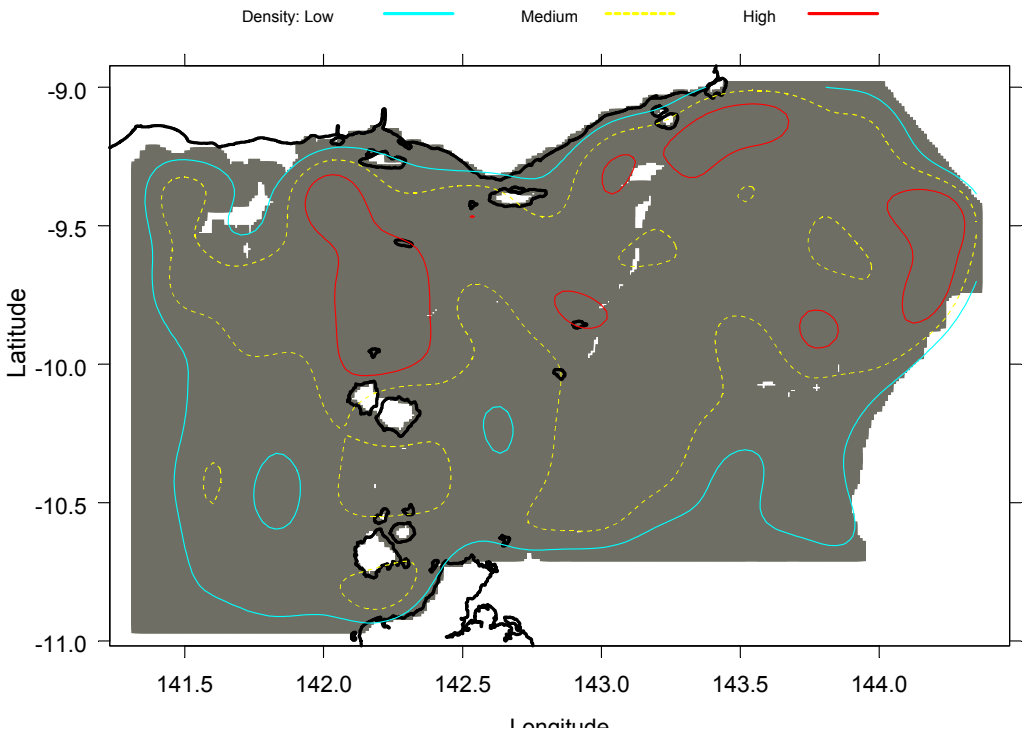


Figure 2.8-12. Average kernel density estimate of the sample sites over 20 independent random samplings. The low, medium and high densities are, respectively, 0.66, 0.88 and 1.10 points per average site separation squared. The 41,285 possible survey sites are indicated by the grey background

Defining trawl substrata

The above has described how 440 substrata were defined from which benthic sampling sites may be chosen. However, somewhat fewer of these same sites (324) were to be selected for trawl sampling and it was necessary to identify which would be the most representative. Although one method would be to simply choose the 324 sites at random, an approach that took advantage of the existing stratification was preferred, to ensure that the selection was heterogeneous. The approach taken was to go back to the superclusters and recompute the number of subclusters required per supercluster to give a total of 324, using the same methodology as before. On average the number of trawl subclusters was about three-quarters (324/440) the number of original subclusters. For instance, supercluster 3, which had 20 original strata, had 15 trawl strata. It was not feasible to try to cluster the sites into trawl subclusters, because there was no way to prevent the original strata straddling several trawl strata. Instead, it was necessary to cluster the sites such that all sites in an original stratum remain together.

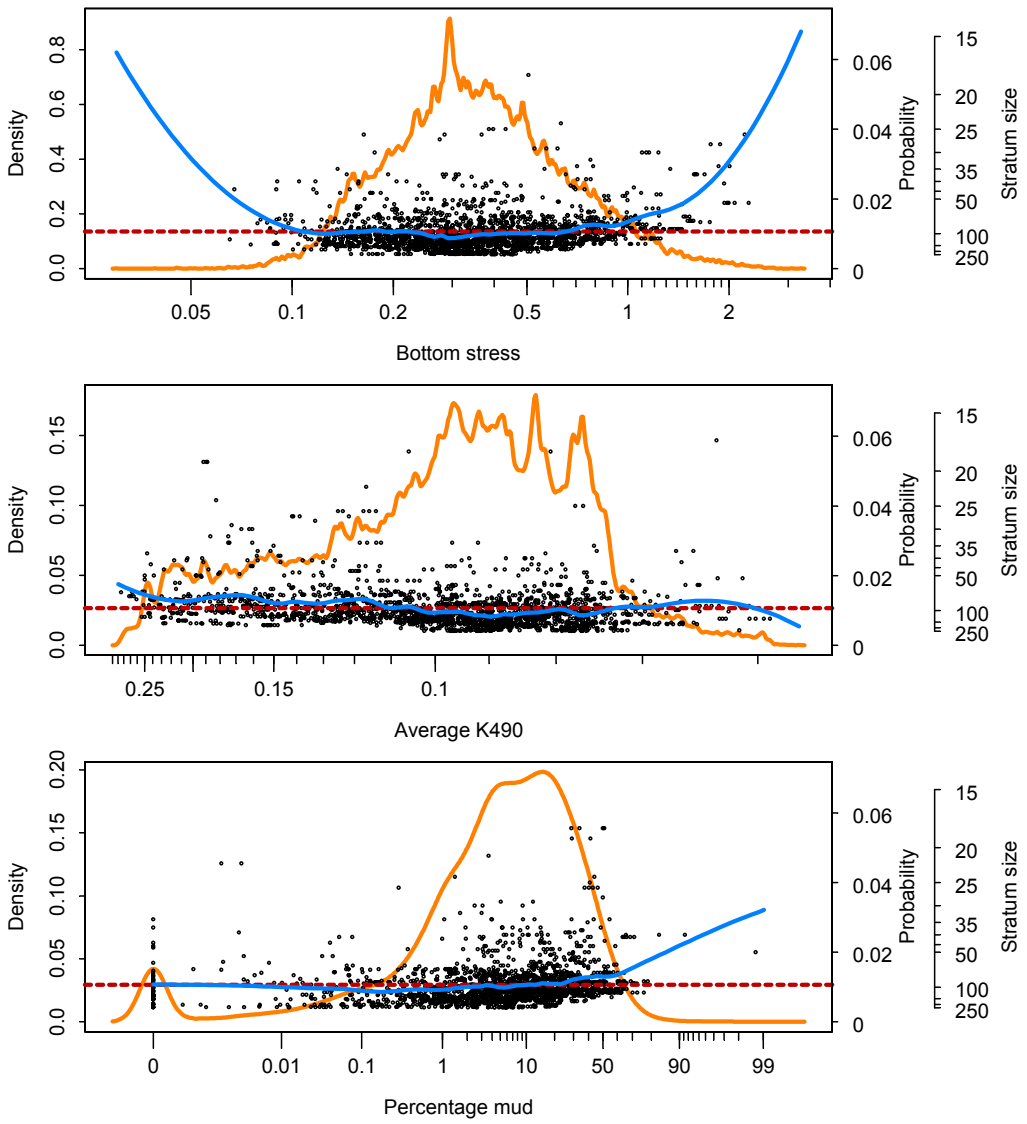


Figure 2.8-13. Local regression smooth of probability of site selection (blue line, right axis) with covariate density (orange line, left axis) for reference. Probabilities for 2000 randomly chosen points (independent of the stratification) are also shown. The dashed horizontal line is the average probability. For clarity we show covariates on a log scale for bottom stress, an inverse scale for K490 and a logit scale for mud.

The simplest way to do this was to cluster the stratum medoids. It was appropriate to use the medoid to represent its stratum as a whole because the medoid lies centrally within the stratum in co-variate space. Since there were at most 20 medoids to cluster, the calculation was computationally trivial. For example, in supercluster 3, substrata 1, 3 and 17 were amalgamated into one trawl cluster, substrata 7 and 15 into a second, substrata 4 and 10 into a third, and substrata 8 and 16 into a fourth, while the other 11 trawl clusters coincide with the original substrata.

After the medoids were clustered, each medoid's trawl substratum number was assigned to all other cells in its substratum. Thus, each cell now belongs to both a substratum and a trawl substratum. Thus for any selection of 440 benthic survey sites, the trawl sites could be selected from these by choosing one from each trawl stratum, either at random or by other objective.

2.8.2. Sample Site Selection

In the previous section, the notion of medoid sampling was raised to illustrate the stratification. Medoid sampling would be a perfectly reasonable method of selecting sites that would deliver the "most typical" cell, with respect to physical covariates, within each of the strata. A random selection of sites from within each of the strata would also be an acceptable method. However, the stratified random method has a relatively high risk of selecting cells too close together and too far apart, creating clumps and voids in the coverage of the region, when in fact a representative coverage that also takes account of the spatial autocorrelation distance was desired. Considering that strata were often fragmented into patches of varying numbers of cells, including single cells, there was also a high risk of selecting isolated cells as sites — these would be less likely to be representative of their stratum due to errors in the covariates. A site selection method that avoided these problems as much as possible was sought.

Initially, a weighted random selection was used, with weights dependent on the spatial geometry of the particular patch within each stratum that the cell belonged to. Cells with fewer neighbours of the same stratum and on the edges of patches (i.e. geographically close to a different stratum) were given less weight, whereas sites in the middle of patches were given more weight. This strategy was intended to reduce the possibility of a site being unrepresentative of its stratum due to errors in the covariates and to avoid selecting adjacent sites. Examination of several weighted random selection options indicated quite a number of adjacent cells being selected and a number of excessively large voids between selected sites. Consequently, a method that more stringently avoided selection of adjacent cells and voids was needed.

The method finally used did not include any deliberate random jittering of site selection. For each of the 440 benthic strata, first all those cells that had the maximum number of neighbours and were the maximum distance from the edge of patches were selected. For many of the strata, several cells met these criteria (total 1698). To remove duplicate cells within strata, the cell with the minimum medoid distance was selected. In about a quarter of cases, the actual medoid cell was selected. This strategy maximized the co-variate representativeness and spatial regularity of the selection, within the desired constraint of the stratification, and minimized the likelihood of clumps and voids, and adjacent, edge and isolated cells.

As described in the previous section, fewer sites could be sampled by trawl methods, so the 440 benthic medoids were clustered to provide 324 most representative options. Of these, 240 were a one to one match with their benthic strata, so no further selection was needed. However, in 84 cases, a trawl site had to be selected from 2-4 benthic site options. In these cases, the benthic site chosen to be sampled by trawl also was, to maintain spatial coverage, that which belonged to the largest patch in its cluster.

The sites selected are mapped in Figure 2.8-14. This site selection process provided a good compromise between coverage of the range of biologically important physical environments in TS and evenness of spatial coverage, given the limited number of sites that could be sampled and the inadequacies of the data available for the stratification. Such a coverage could not be achieved with regular grid sampling or completely randomised sampling.

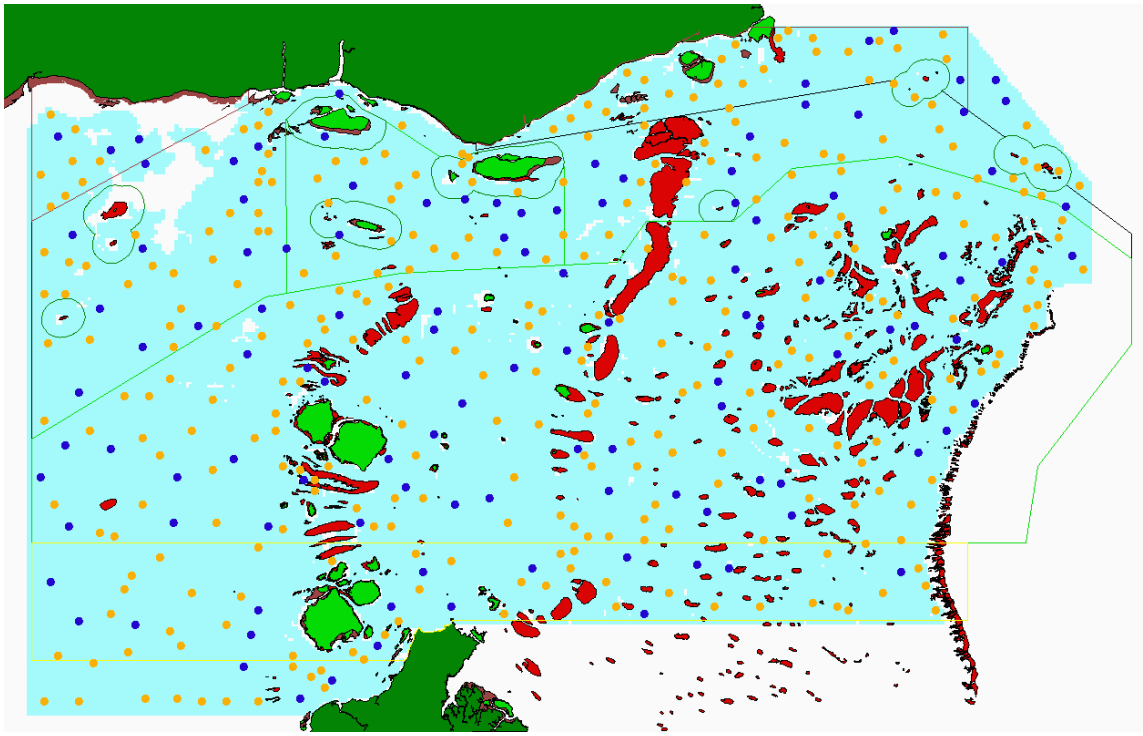


Figure 2.8-14 Map of the sites selected for sampling the seabed in Torres Strait, overlaid on a background of all cells included for possible selection (light blue). White areas were excluded as outside the study area or too shallow for navigation. ●: sites for benthic and trawl sampling, ●: sites for benthic sampling only.

2.8.3. Mapping the Physical Characterization

The biologically informed stratification developed in section 2.8.1 is a physical characterisation of Torres Strait that can be considered an *a priori* surrogate for patterns in seabed biodiversity assemblages, to be tested and improved by the future sampling to be conducted by the CRC-TS Seabed Mapping Project. Given the likely interim utility of this information, a method of representing this complex multi-variate data in a single map was sought.

2.8.3.1. The Colour Key

The objective was to produce a map of the Torres Strait with similar colours representing similar physical environments, which could be expected to have similar benthic biotic assemblages. The colour mapping should encompass as much information as possible in a reduced form — this was achieved by deriving a colour key from the first and second principal components of the biological importance weighted covariate data used in the stratification. A biplot of the principal components and physical variable vectors would provide a key to the environmental characteristics of the map. Particular directions in the biplot that corresponded to important covariates, should be coloured in an intuitive manner. Red was used to denote high bottom stress and green to denote high average

chlorophyll A (which correlated with K490). Blue corresponded with depth. High density areas of the biplot (common areas) should have a neutral colour such as white or grey.

A further desirable property of the colour key is that it should cover the data space compactly, to avoid large areas of the key having no data and wasting part of the colour space. The colour key should therefore be shaped to conform to the distribution of the data in principal components (PC-)space. This was done by mapping a circular colour disk to a simply connected region enclosing the data. In order to do this, it was necessary to first define a boundary of the data in PC-space. One way to do this was to find the convex hull; however, for the Torres Strait data, this included a void in which no data existed. Instead, a more compact boundary was found by computing a two-dimensional kernel density function and delineating a contour of sufficiently low density. The boundary is partly concave.

Having defined a boundary, there were two alternative methods for mapping the colour disk to the region inside the outer density contour boundary: polynomial mapping and conformal mapping. The polynomial mapping was found to be more flexible but because of the partly concave shape there was not always a one-to-one mapping between PC-space and colour space, and it was non-trivial to invert from PC-space to colour.

2.8.3.2. Conformal Mapping

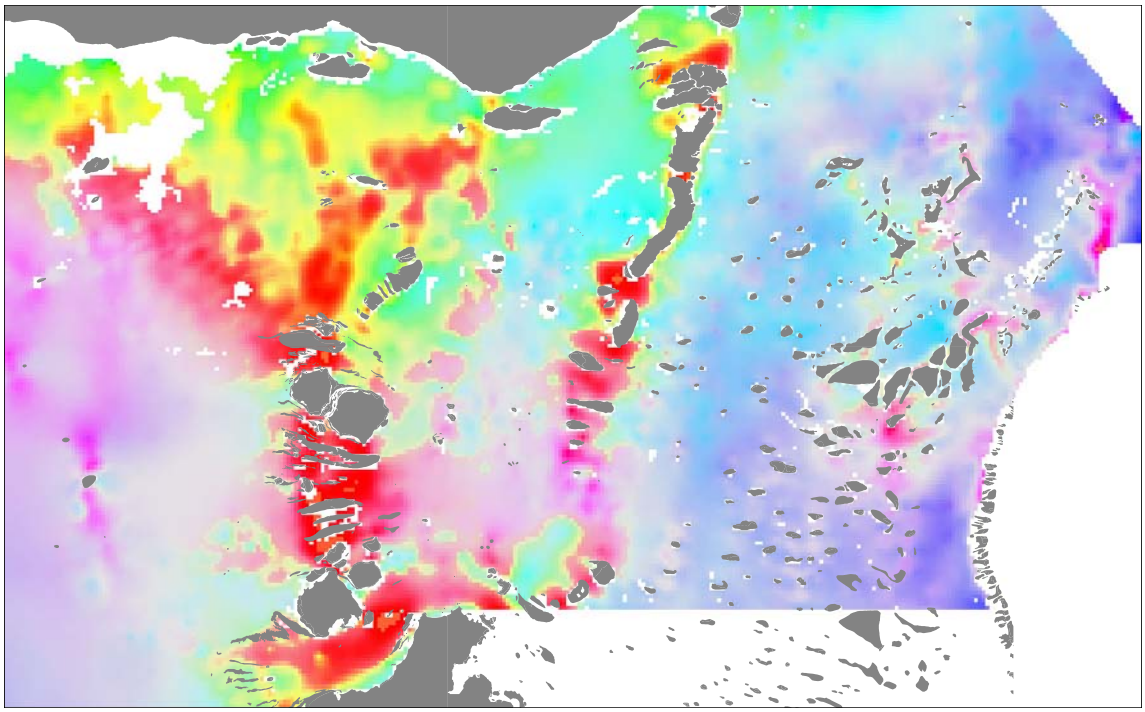
The conformal mapping method originates from complex number theory. A mapping from the colour disk to a simply connected polygon is expressible as a complex integral, whose parameters must be estimated by a non-linear algorithm. Trefethen (1980) provided a FORTRAN program to compute this integral. An interface to this code was developed that runs in R. Conformal mappings have certain benefits (such as local preservation of angles) but most importantly they are guaranteed to map the interior of the colour disk to the interior of the polygon (i.e. the mapping will not stray outside the boundary).

As with the polynomial method, the point in PC-space that the centre of the disk was mapped to was specified. The matching of points on the edge of the disk with vertices of the polygon was done by the non-linear algorithm. In order to match intuitive colours to the desired directions in PC space, it was necessary to impose a further transformation on the colour disk, which amounted to an angular stretch and shift. This was done using a periodic piecewise linear function of the angle. To complete the physical characterisation map, each grid cell must have a colour associated with it. Hence, the colour key mapping must be inverted, so that points in PC-space become mapped to points in colour space. This inverse mapping is available in FORTRAN code (Trefethen, 1980).

The resulting physical characterisation map of Torres Strait is shown in Figure 2.8-15. High bottom stress areas were coloured red, high Chlorophyll/K490 areas green, and the mud direction was coloured cyan. Sites coloured cyan have high levels of mud. Deeper areas tend to be blue.

The colouring of a map to highlight different covariates can be highly effective at illustrating similar and different physical environements, especially when the colour space has been fully utilized. The two colour mapping techniques investigated each had advantages and disadvantages. The main disadvantage of the polynomial method was discussed above. On the other hand, the conformal mapping method tended to cover jutting-out parts of the PC-space from fairly small regions in colour

space (e.g. the red area of the key in Figure 2.8-15). This would be a significant disadvantage if such an area were densely populated with data.



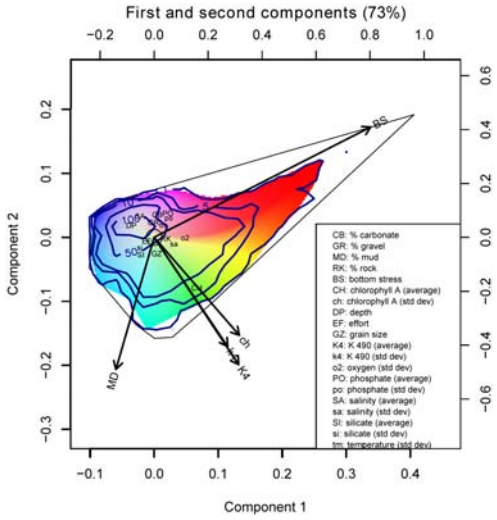


Figure 2.8-15 Map of the stratification of the Torres Strait seabed, with similar colours representing physically similar strata. *Inset right*: colour key from a bi-plot of the first and second principal components of the biologically weighted physical data. Contours (blue lines) of a kernel density estimate of all 41285 scores is also shown at levels 1, 5, 10, 50 and 100. The convex hull of these scores is shown by a thin black line. Loadings of important variables are denoted by black arrows and are labelled. The less important variables (those with smaller loadings) are denoted by small labels. These two components explain 73% of the variance. In the background is the colour key, which is a mathematically distorted colour disk mapped to the interior of the outermost contour using conformal mapping. Red has been chosen to align with high bottom stress and green with high K490 and chlorophyll A. average. Cyan aligns with mud. Common cells close to the mode of the density function have less saturated colours (grey).