MARINE MONITORING PROGRAM

Modelling the environmental drivers and abundance of seagrass communities in Cleveland Bay

CSIRO DATA 61

Australian Government
Great Barrier Reef
Marine Park Authority

Published by the Great Barrier Reef Marine Park Authority

ISBN 9780648357186

A catalogue record for this publication is available from the National Library of Australia

This publication should be cited as:

Lawrence, E., 2019, *Modelling the environmental drivers and abundance of seagrass communities in Cleveland Bay*. Great Barrier Reef Marine Park Authority, Townsville. 56pp.

Front cover photo: Monitoring seagrass at Cape Cleveland. © Great Barrier Reef Marine Park Authority

DISCLAIMER

While reasonable efforts have been made to ensure that the contents of this document are factually correct, the Commonwealth of Australia, represented by the Great Barrier Reef Marine Park Authority, does not accept responsibility for the accuracy or completeness of the contents, and shall not be liable for any loss or damage that may be occasioned directly or indirectly through the use of, or reliance on, the contents of this publication.

Comments and questions regarding this document are welcome and should be addressed to:

Australian Government
Great Barrier Reef
Marine Park Authority

Great Barrier Reef Marine Park Authority
280 Flinders St Townsville | PO Box 1379 Townsville QLD 4810
Phone: 07 4750 0700
Fax: 07 4772 6093
Email: info@gbrmpa.gov.au
www.gbrmpa.gov.au

This project was supported by the Great Barrier Reef Marine Park Authority through its Marine Monitoring Program, and with funding from the Australian Government Reef Program, and the Australian and Queensland governments Reef 2050 Integrated Monitoring and Reporting Program.

Table of Contents

Figures

Acknowledgments

I thank Katherine Martin and Carol Honchin from the Great Barrier Reef Marine Park Authority, for guidance and feedback during development and delivery of the project. The work was based on a large number of discussions with Catherine Collier, Alexandra Carter, Michael Rasheed and Len McKenzie (TropWATER) and I thank them for their advice and contributions to the technical methodology.

I thank the Port of Townsville Limited for their support and funding of the long-term seagrass monitoring program for Cleveland Bay. I would like to thank the many James Cook University and Fisheries Queensland staff who have contributed to the monitoring program over the years.

Executive summary

This report presents statistical analysis of seagrass data from monitoring in Cleveland Bay. Its objective was to determine which environmental drivers best explain where seagrass species communities are found, and what drives the trends in their abundance.

Seagrass biomass data from Cleveland Bay (located near Townsville, Australia, in the Great Barrier Reef Marine Park) was analysed using multivariate regression trees to establish the seagrass community types in Cleveland Bay, and to identify environmental drivers of the types. This analysis forms part of a broader project on deriving ecologically relevant load targets to meet desired ecosystem condition, being conducted under the National Environmental Science Programme.

The multivariate regression tree analysis indicated that there are up to nine community types in Cleveland Bay. The environmental drivers of the different community types are:

- relative exposure (indicates the spatial extent of intertidal substratum exposed at percentile intervals of the observed tidal range)
- sediment type
- water type
- depth

While relative exposure, sediment type and water type are readily available at a national scale, sediment type is not routinely collected. Sediment type was important in differentiating community type in many of the tree analyses undertaken. The routine collection of sediment type would improve the predictability of community type in areas where little biological data is available.

In the intertidal region, the community assemblages primarily vary by the amount of *Cymodocea*, *Halodule* and/or *Zostera* present. In the subtidal region, the assemblages are similar but with increasing depth show overall higher biomass values for *Halodule* and *Halophila spinulosa*.

Once the species communities were established using the multivariate regression trees, the total biomass in each tree node was modelled to determine any trends in abundance in the different community types. Over the study period,from 2007—2016, the mean biomass of seagrass fluctuated showing significant loss and subsequent recovery in all of the seagrass communities. While the trends among communities were similar, the overall ‘usual’ or recovered state varied. There was sufficient data in most of the communities in most of the study years to estimate mean absolute abundance and associated uncertainty. Through further examination of these absolute abundance estimates, the broader project is expected to be able to set desired state targets for seagrass for the community types in the bay.

The methodology used is robust and could be used in other areas of the Great Barrier Reef where sufficient data exists.

Introduction

CSIRO performed analysis of seagrass biomass data to assist authors of the “Deriving ecologically relevant load targets to meet desired ecosystem condition for the Great Barrier Reef: a case study for seagrass meadows in the Burdekin region” National Environmental Science Programme (NESP) project.

The NESP project aimed to develop seagrass condition targets for habitats of the Great Barrier Reef as benchmarks, against which to report on ecosystem health, at a case study location in Cleveland Bay near Townsville. This location was chosen because of the long-term community composition and biomass data, sampled intensively in both space and time (Figure 1).

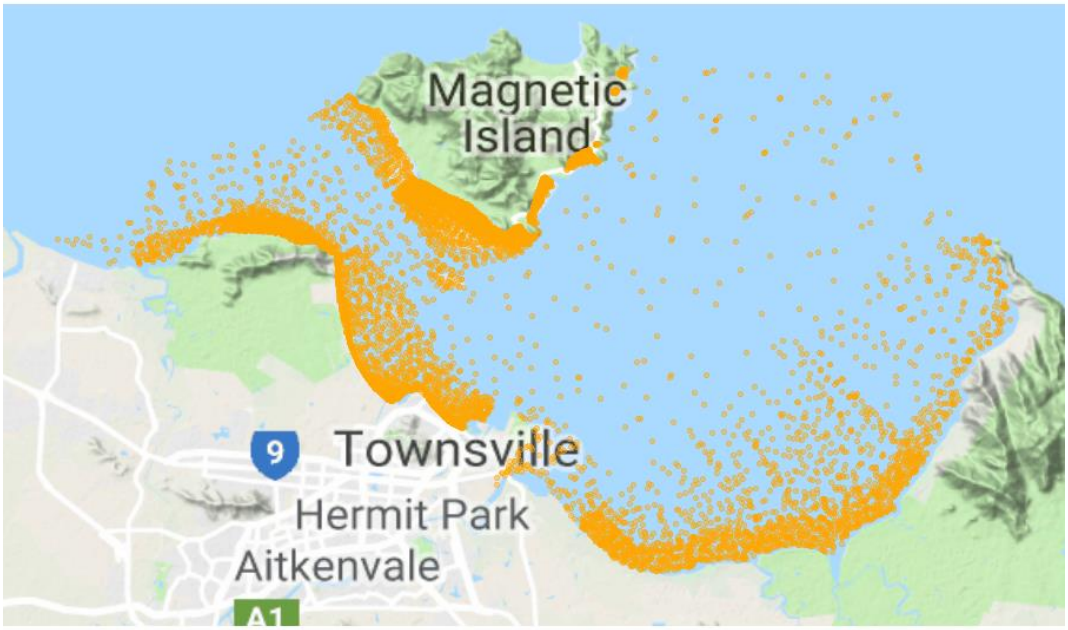

Figure 1: Location of sites where seagrass biomass has been estimated (2007-2016)

The broader project team aimed to develop a framework for setting and reporting seagrass condition targets (‘desired-state’) in each of the habitat types in Cleveland Bay, considering both seagrass state and seagrass trajectory. While it was not CSIRO’s task to come up with desired state as such, our analysis will help the investigators to determine appropriate targets and frameworks in this ‘data-rich’ area. For this reason, we concentrate on the statistical interpretation of the analysis and encourage the reader to refer to the final NESP project report (to be published in 2020) for a more specific biological interpretation.

We first outline the data available, followed by the two components of the analysis:

1) Establishing the seagrass community types in Cleveland Bay: Seagrass meadows are dynamic. The seagrass community species composition varies throughout Cleveland Bay. Different community types might be associated with varying degrees of abundance under typical or good environmental conditions. It is important to distinguish between the community types to better understand what desired state condition targets might be appropriate in different parts of the Bay.

2) Analysing the temporal trends in those communities: the temporal trends over the last ten years provide an indication of the different states that seagrass might exhibit.

Data

The biological data used for this analysis is made up of almost 8000 observations collected from 2007 to 2016. The dataset is primarily comprised of the long-term ports monitoring and assessment data (Wells and Rasheed 2017) supplemented by data collected under other projects with similar sampling protocols (Carter et al. 2016; Davies and Rasheed 2016). The distribution of observations covers much of Cleveland Bay (Figure 1).

The biological data on species and their abundance was combined with available habitat characteristic and environmental pressures to create a dataset with the following key variables:

- Biomass (grams dry weight m^2) estimated using a calibrated visual estimation technique (summarised in Wells and Rasheed, 2017) for each species at each site. The species recorded *are:*
    - *Cymodocea serrulata* (CS)
    - *Halophila decipiens* (HD)
    - *Halodule uninervis* (HU)
    - *Halophila ovalis* (HO)
    - *Halophila spinulosa* (HS)
    - *Zostera muelleri* subsp. *capricorni* (ZC).
- Latitude and longitude (collected using GPS in the field)
- Sediment2: derived sediment (the sediment type recorded in the field aggregated to the categories: Coarse Sand, Mud, Reef, Rock, Rubble, Sand).
- ITEM_REL_m: the extraction from the relative exposure raster where:
    - 0 is never exposed
    - 1-9 is exposed at increasing amounts of time (see Carter et al. 2018)
- depth_final: depth (m) (see Carter et al., 2018)
- Habitat_1: habitat classification developed in Carter et al. 2018, a spatially-explicit habitat classification scheme developed for the entire Great Barrier Reef based on water depth and water quality. The categories relevant to Cleveland Bay are:
    - Coastal intertidal
    - Coastal shallow subtidal
    - Deep subtidal.
- TSV_compos: an indicator describing whether the data was collected as part of the ports long-term monitoring program (1: long-term monitoring, 0: otherwise)
- WT2: number of weeks in the previous wet-season that water type was exposed to ‘primary’ waters (categories 1-4, Devlin et al. 2017)

While other environmental covariates (primarily climate data) were compiled for this analysis the resolution was too coarse to be useful (all observations have the same value within a given year).

Part I: Regression trees define communities

Methods

Regression trees (Brieman et al. 1984) are a machine-learning method for constructing prediction models from data. The hierarchical nature of regression trees means that the response to one predictor variable depends on others higher in the tree, thereby modelling interactions. Trees do not include *a priori* assumptions about the relationships between the response and predictor variables, allowing for non-linear relationships between variables.

Regression trees explain the variation of a continuous response variable based on predictor variables by partitioning data into mutually exclusive groups, where observations in a particular group are as homogeneous as possible (Brieman et al. 1984, Clark and Pregibon 1992). Starting with all data in a single node at the top of the tree, each split of the tree is designed to minimize the total sum of squares (the impurity measure) of the response variable about the mean of the node, within the two nodes formed by the split. The splitting procedure is then applied to each node separately and the process continued until some stopping rule is reached. Generally, trees are grown to their maximal size (over fitted), stopping only when all the terminal nodes are homogenous or have reached a pre-specified number of observations or nodes. The terminal nodes ('leaves' of the tree) represent groups of observations, or clusters, formed by the tree. Depending on the objectives of the analysis the tree is then pruned to an appropriate size, where size refers to the number of terminal nodes.

The fit of a tree is usually defined by the relative error (RE), the total impurity of the leaves divided by the impurity of all the data combined (the root node). However, RE gives an overoptimistic estimate of how accurately a tree will predict to new data. The predictive accuracy is better estimated by the cross validated relative error (CVRE; Hastie et al. 2009).

We fitted regression trees to each single species separately to determine the environmental drivers that may explain the differences in the abundance of seagrass at each site. We fitted the models in the rpart package (Therneau et al. 2017) in R 3.4.3 (R Core, 2017). As we are essentially looking to cluster the sites spatially (not predict the abundance at a site at a point in time), we did not use any temporal variables in our models. Instead we are just planning to roughly categorise where each species of seagrass is found, on average, and determine which variables might be driving the distribution of abundance values.

The environmental variables we used as covariates were:

- Habitat_1
- Sediment2
- depth_final
- ITEM_REL_m
- WT2

During our initial analyses we also included latitude and longitude in our regression tree's, however we later decided to remove these variables to allow other important environmental gradients to drive the results.

We performed 10-fold cross validation and used the "1se" rule (De'ath 2002) to reduce overfitting and set the 'minbucket' (minimum number of observations in a terminal node) to 50 to ensure that no clusters were overly small (a minimum number of observations will be needed in each cluster for later temporal analysis). Using the tree for each species we then predicted which cluster each site would fall into and coloured each cluster differently to show how the clusters were spatially separated.

While the single species regression trees were informative at a species level, this task is focussed on identifying different seagrass community types (or species assemblages). The individual species regression trees resulted in different "splits" or clusters for each species, so we moved to Multivariate Regression Trees (MRTs). MRTs are capable of simultaneously estimating the mean response of multiple dependent variables (Larsen and Speckman 2004). Sums of Squares Multivariate Regression Trees (SS-MRTs; De'ath 2004) are an extension of regression trees and minimize the total sum of squares over the multivariate response. The total sum of squares (impurity measure) is defined as $\sum_{i,j}(x_{ij} - \bar{x}_j)^2$, where x_{ij} is the transformed biomass of species *j* recorded at site *i, and* $\bar{x}_j$ is the corresponding mean value across sites in a particular tree node.

MRTs can be used to describe and predict relationships between multiple species and environmental variables (De'ath 2002). Each cluster represents a species assemblage, and its environmental values define its associated habitat. Using this method to identify habitats associated with each species assemblage, allows us to later predict community types in areas where there is environmental data but little to no seagrass composition data.

We fitted an MRT to the transformed (square root) biomass of each species of seagrass in the form of a multivariate response. We used the square root transformation (and tested for the sensitivity of this transformation) so that the more dominant species did not entirely drive the results. The environmental covariates were the same as those used for the single species models. We completed our analysis using the mvpart package (De'ath 2014) in R (Note: that this package is available in archive form on CRAN https://cran.r-project.org).

We considered removing the zero biomass values and subsequently only clustering sites where seagrass was present. This almost halved the dataset. However, given the spatial overlap between the zero and non-zero biomass records (Figure 2) we do not think it is justified to remove the zeros from the analysis, nor necessary, given we are trying to cluster areas that are alike (rather than reliably predict the biomass estimate of a particular species). Figure 2 shows that there is evidence of the presence of seagrass through most of Cleveland Bay, at some point over the ten year study period.

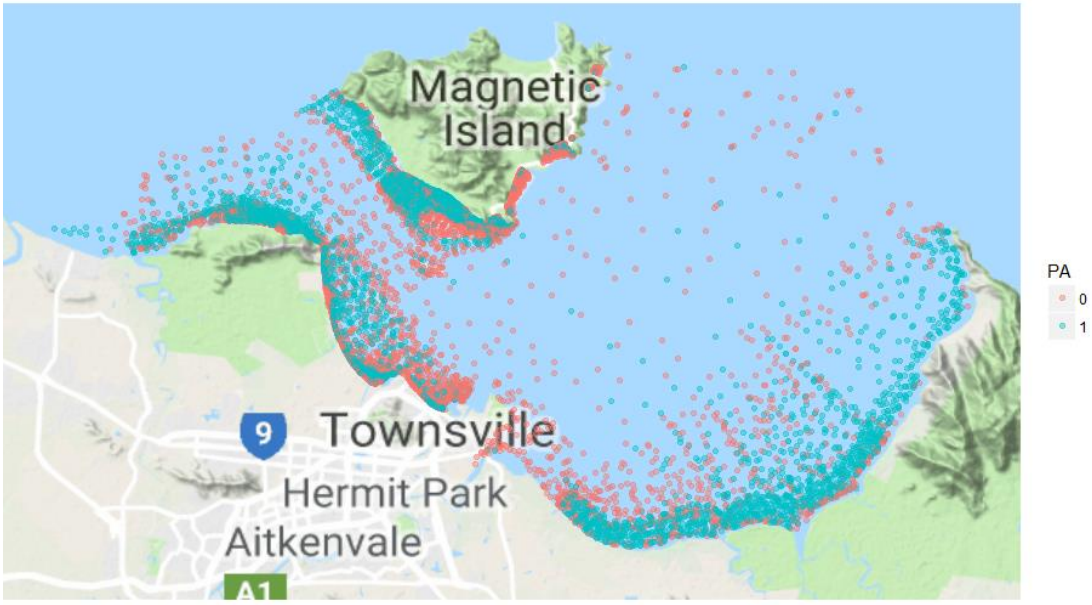

Figure 2: Distribution of zero (0: pink)/non-zero (1: blue) biomass estimates on combined data 2007-2016

We went through a process of exploratory analyses and sensitivity testing of the models to determine our final set of results, remembering that the goal of this step was to classify observations into similar community types. The analyses we ran were (using the model parameters previously described, unless noted):

1. A single MRT (Base MRT) with all data and all relevant and available environmental variables. The response variable was a square-root transformed biomass matrix of all species.
2. The Base MRT where the data was restricted to just the ports long-term monitoring data. The data not collected as part of the long-term monitoring was more sporadic spatially and collected over a shorter time period. We wanted to ensure that this was not driving the analysis, given the less consistent nature.
3. Separate MRTs (equivalent to Base MRT model) for Intertidal and Subtidal habitat types. The reason for splitting these was to ensure that the community types did not "cross-over" into different proposed Reef Integrated Modelling and Reporting Program's habitat categories (see Carter et al. 2018; Udy et al 2018).
4. Separate MRTs (equivalent to Base MRT model) for each year. This was to test the sensitivity of the community classifications to "good" and "bad" years. As these analyses were based on much smaller datasets, we reduced the 'minbucket' parameter to 20.
5. Based on the separate yearly MRTs, removed all low biomass years (2009-2012) and ran a combined "good" year analysis (equivalent to Base MRT). Removing these years ensures that we are quantifying what the communities look like in what might be considered more 'desirable' years.
6. Separate regression trees for Intertidal and Subtidal based on presence-absence data. Regression trees can be sensitive to the transformation of the response variable. In addition, usually the higher biomass species have greater influence on the analysis. For this stage of the analysis we opted to give all species an equal weight by basing the analysis on the presence-absence data.
7. Estimated the abundance of each community type based on the nodes determined by the MRTs on the presence-absence Intertidal and Subtidal data excluding 2009-2012.

Results

While we fitted regression trees to individual species, we have put these results in Appendix A as each species resulted in a different series of splits. While interesting, the differences in splits mean that aggregation to community level would not be straight-forward. Rather than trying to develop clusters based on individual species' analyses we opted to use a multivariate response (all species) as described in the methods.

Regression trees are interpreted by working your way down the tree and examining the splits. For example, if the first split is labelled as WT2<20.5 and WT2≥ 20.5 (Figure 3) the data is first divided based on the WT2 variable, with observations having values less than 20.5 going to the left of the tree and those greater than or equal to, to the right. For our data this means that if the water type was 'primary' for more than 20 weeks (WT2≥20.5), there may be some noticeable change in the biological response/community type. Once the tree has finished splitting the data, we refer to the bottom 'leaves' of the tree as nodes or clusters. The number to the left of the node is the average biomass value across all species at that node (which isn't very relevant here) and the number to the right is the number of observations that fall into that node. The histogram shows the distribution of square-root transformed biomass values for each species in that node. By comparing these distributions across nodes, we get a good idea about how the community types differ based on the environmental drivers (variables causing the splits in the tree). The CV Error is the cross-validated relative error and is the best indication of the model fit here. Note that we expect the CVRE values to be very high as we are effectively trying to model seven species at the same time.

We start by showing the results of the Base MRT (Figure 3) and then work through the other sensitivity analyses we performed.

1. Base Multivariate Regression Trees with all data

The Base MRT has four terminal nodes with most of the observations (5417) falling in the node to the far left of the tree, where the square-root transformed biomass for all species is almost zero except for a small amount of *Halodule uninervis*. The second node has slightly more *Cymodocea serrulata, Halodule uninervis* and *Zostera muelleri*. The third has more Zostera *muelleri* but less of the other species than Node 2 and the fourth has far more *Zostera muelleri*. So an interesting interpretation of this tree, when comparing Node 3 to Node 4, would be that when the sediment type is mud the biomass of *Zostera* is higher.

Figure 4 shows how the nodes of the Base MRT are represented spatially. Each node is coloured differently, with the nodes of the tree increasing in number from left to right. For example, the far left is Node 2, followed by 4, then 6 then 7 (the numbering is automatically given by the algorithm based on the MRT table so may not seem intuitive). Node 2 represents the bulk of the data and is fairly contiguous, with the others being smaller and having less obvious spatial boundaries.

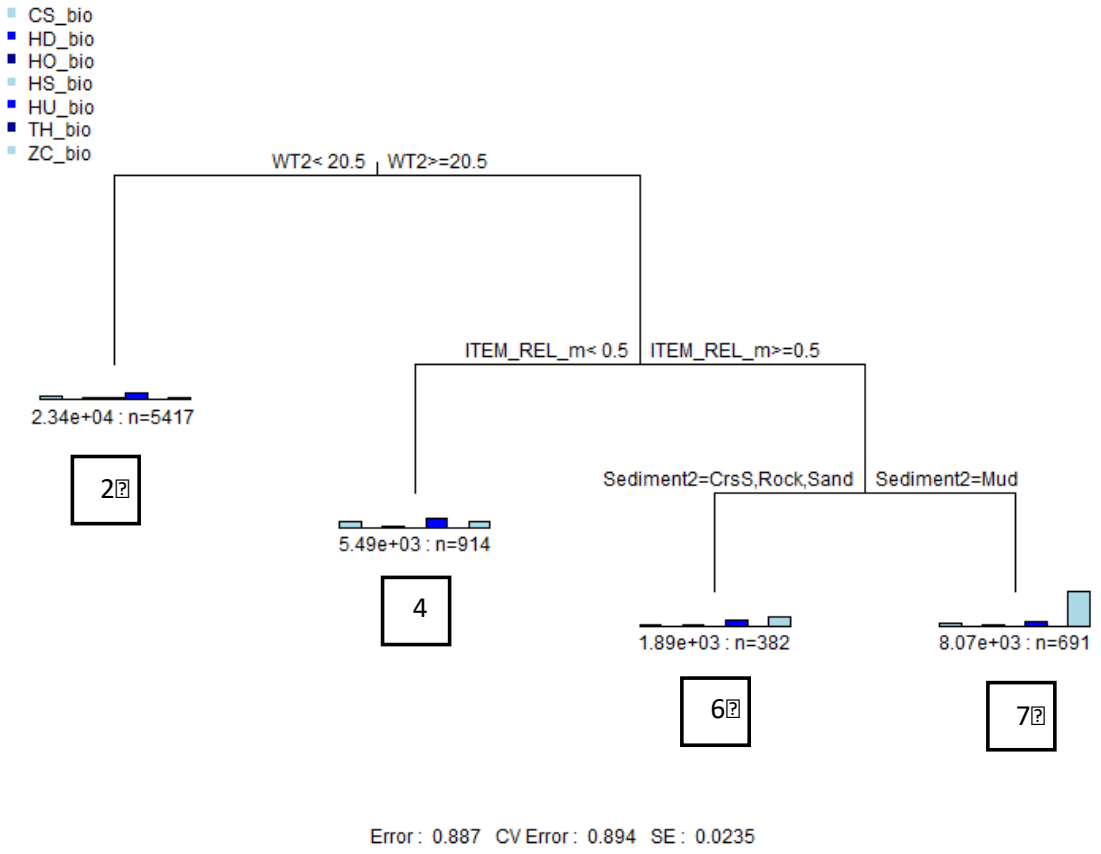

Figure 3: MRT on all data where the response matrix is square-root biomass (Base MRT). The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 4.

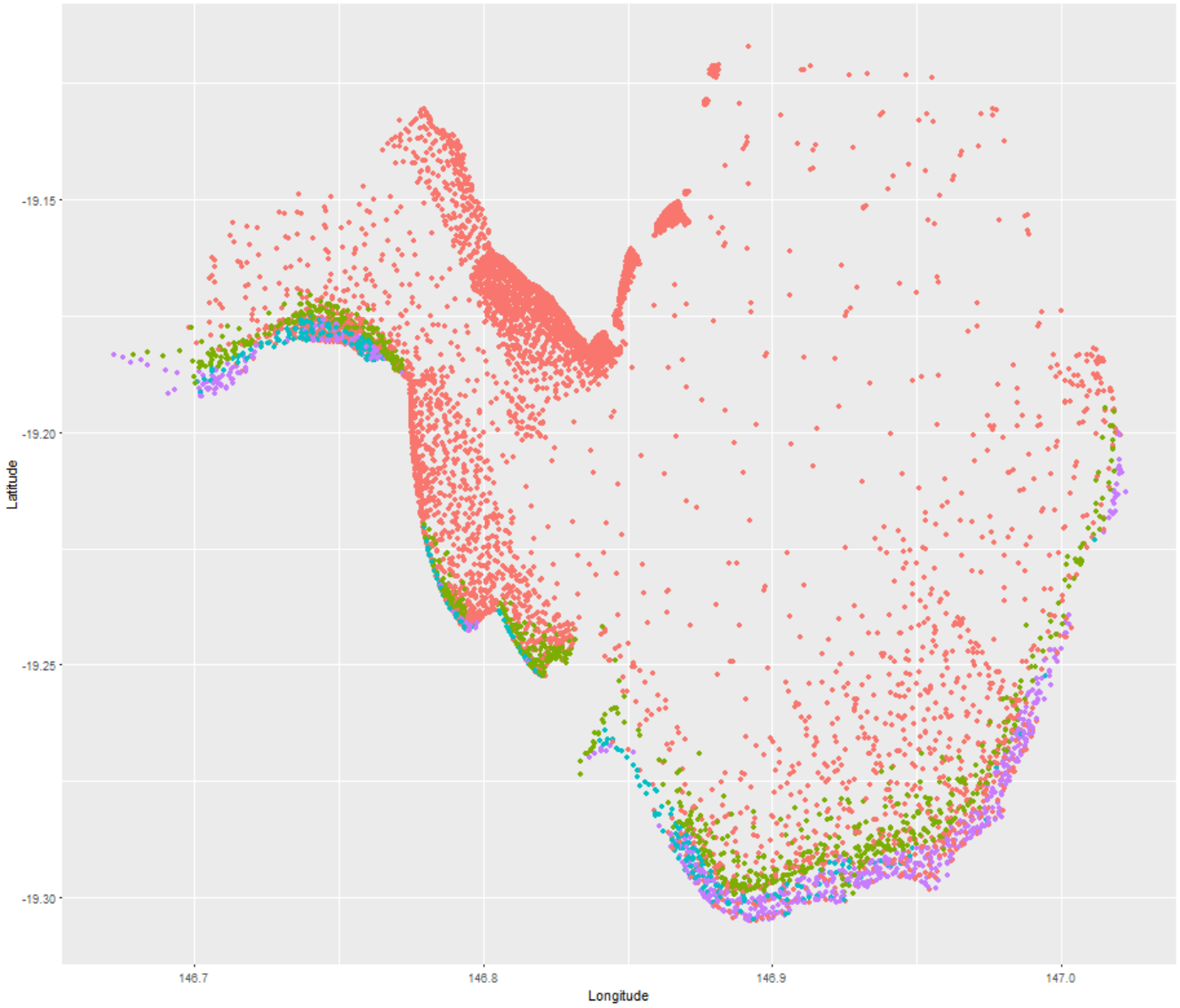

Figure 4: Spatial distribution of node membership for all sites classified using the Base MRT on square-root transformed biomass. Colour code for nodes: ● = 2, ● = 4, ● = 6, ● = 7.

2. MRT restricting the data to the long-term monitoring data

We fitted an MRT restricting the data to the long-term monitoring data (Figure 5 and Figure 6). The variables used to split the tree are similar to the Base MRT (with the addition of depth) and the splits are the same for both water type and sediment type. The relative exposure cut-off has moved from 0.5 to 1.5. Given the similarity between these splits and those in the Base-MRT, we can see no justification for basing the analysis only on the long-term monitoring data and so the remainder of the analyses use the data in entirety.

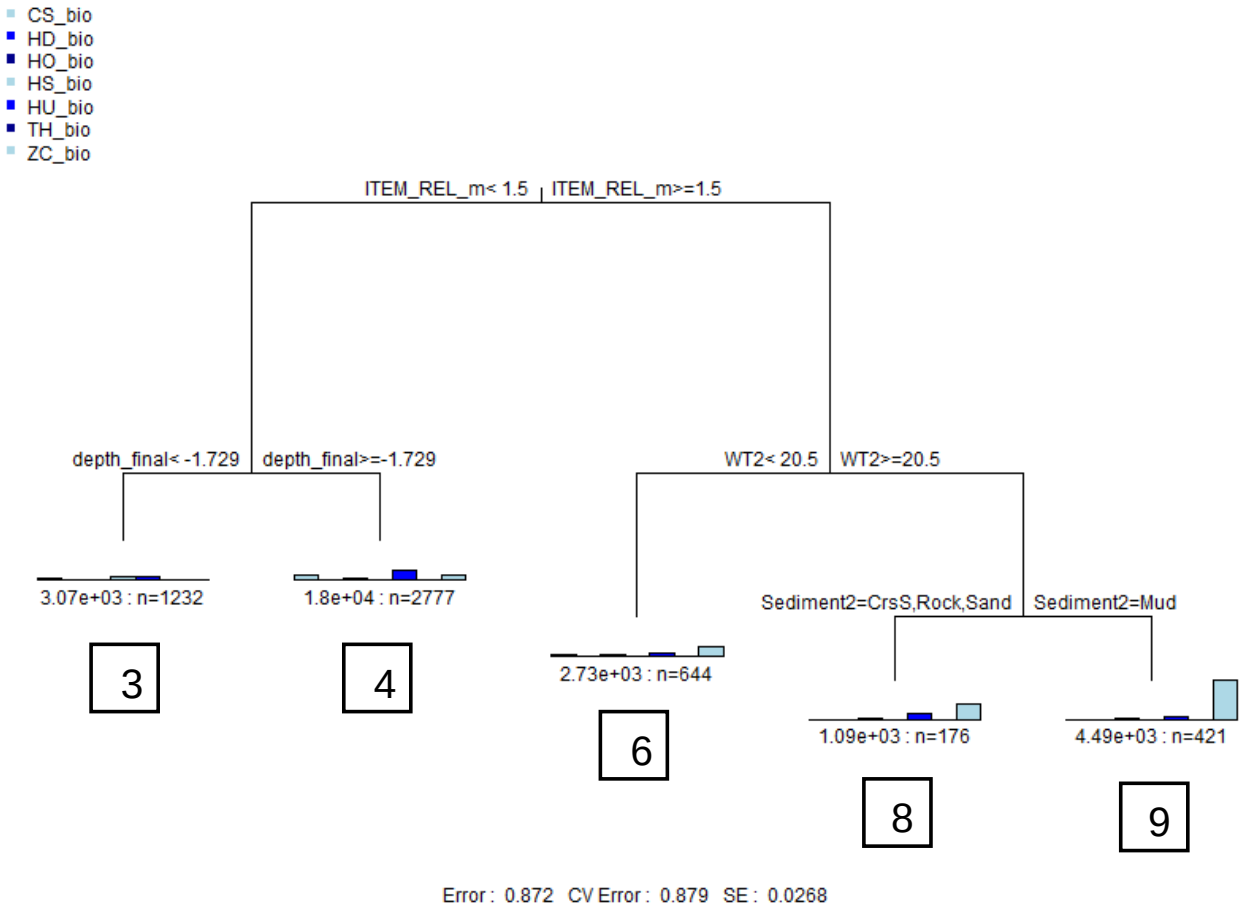

Figure 5: MRT on all long-term monitoring data where the response matrix is square-root transformed biomass.The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 6.

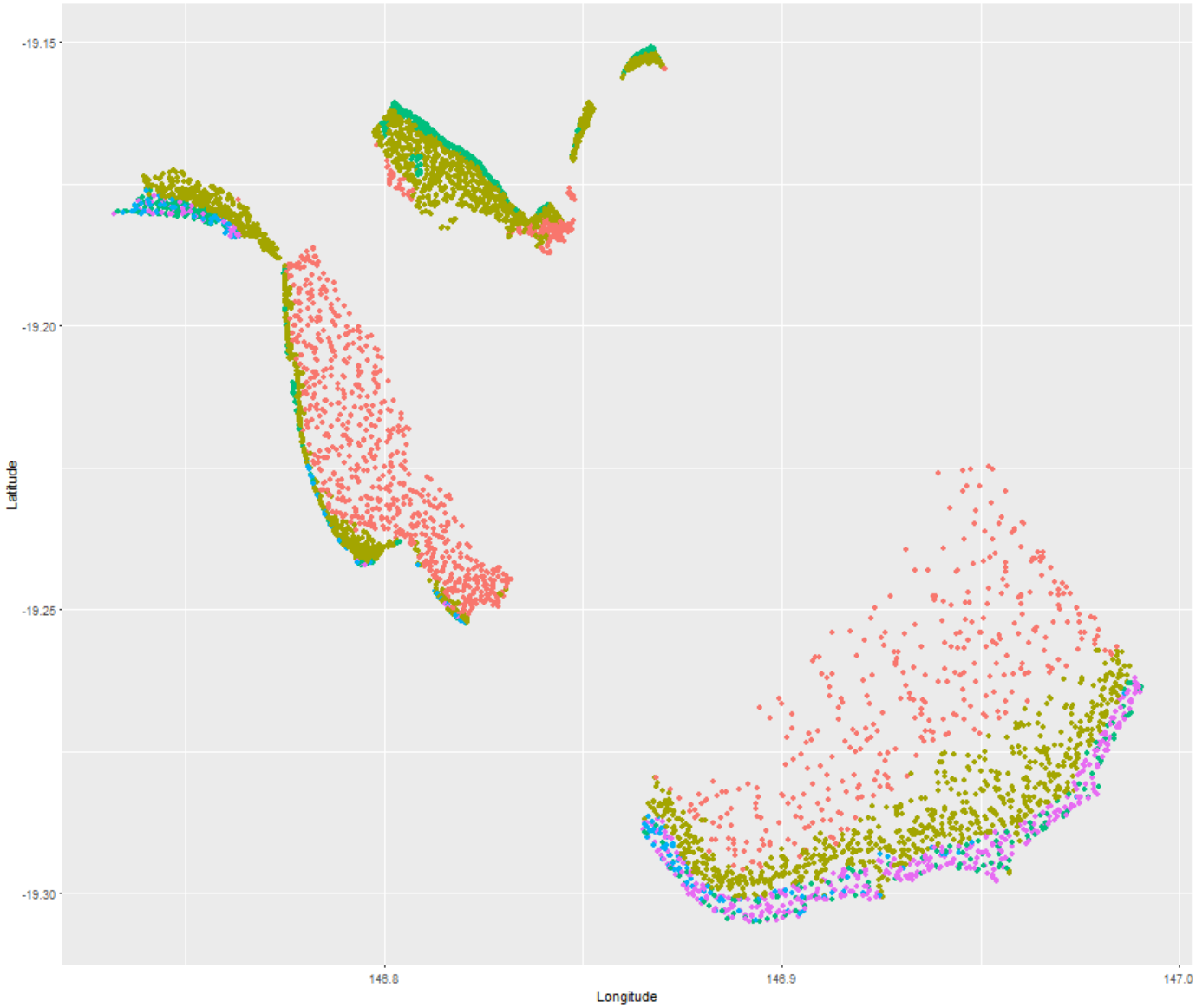

Figure 6: Spatial distribution of node membership for long-term monitoring sites classified using the MRT on square-root transformed biomass. Colour code for nodes: ● = 3, ● = 4, ● = 6, ● = 8, ● = 9.

3a. MRT where the habitat is coastal intertidal

Sub-setting the data to where the habitat is coastal intertidal (Figure 7) resulted in the same variable splits as the Base-MRT (Figure 3) just in a different order (sediment followed by relative exposure (ITEM_REL_M) rather than the other way around). For this reason, Figure 8 looks similar to Figure 4 with just the subtidal sites not mapped.

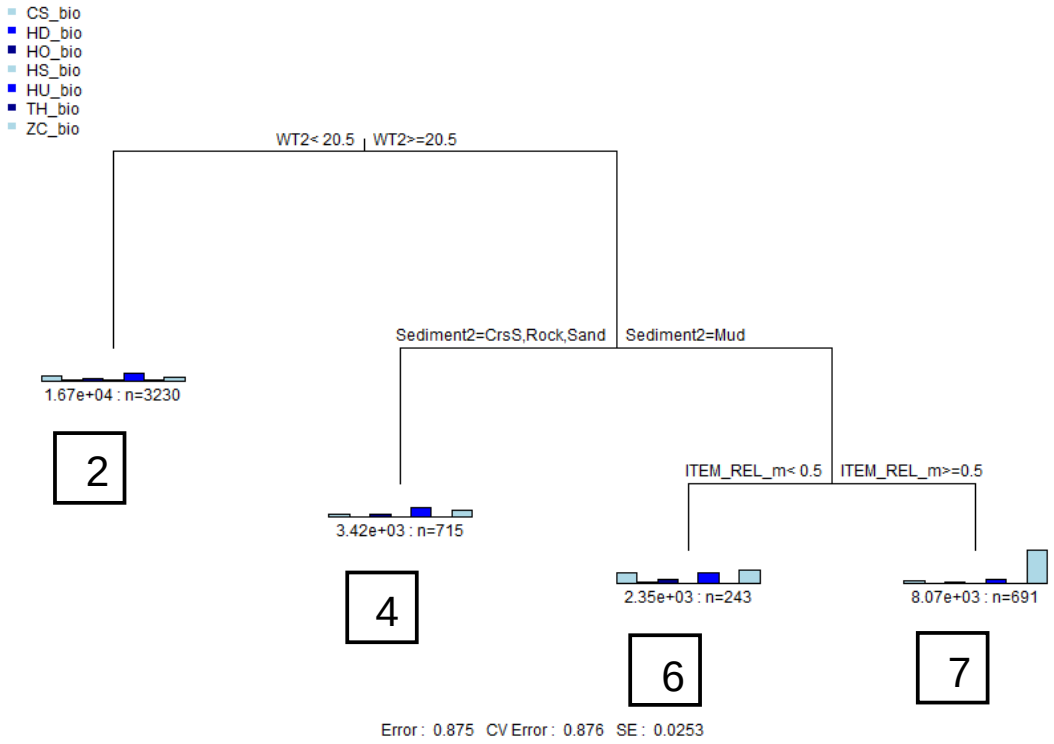

Figure 7: MRT on coastal intertidal data where the response matrix is square-root transformed biomass. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 8.

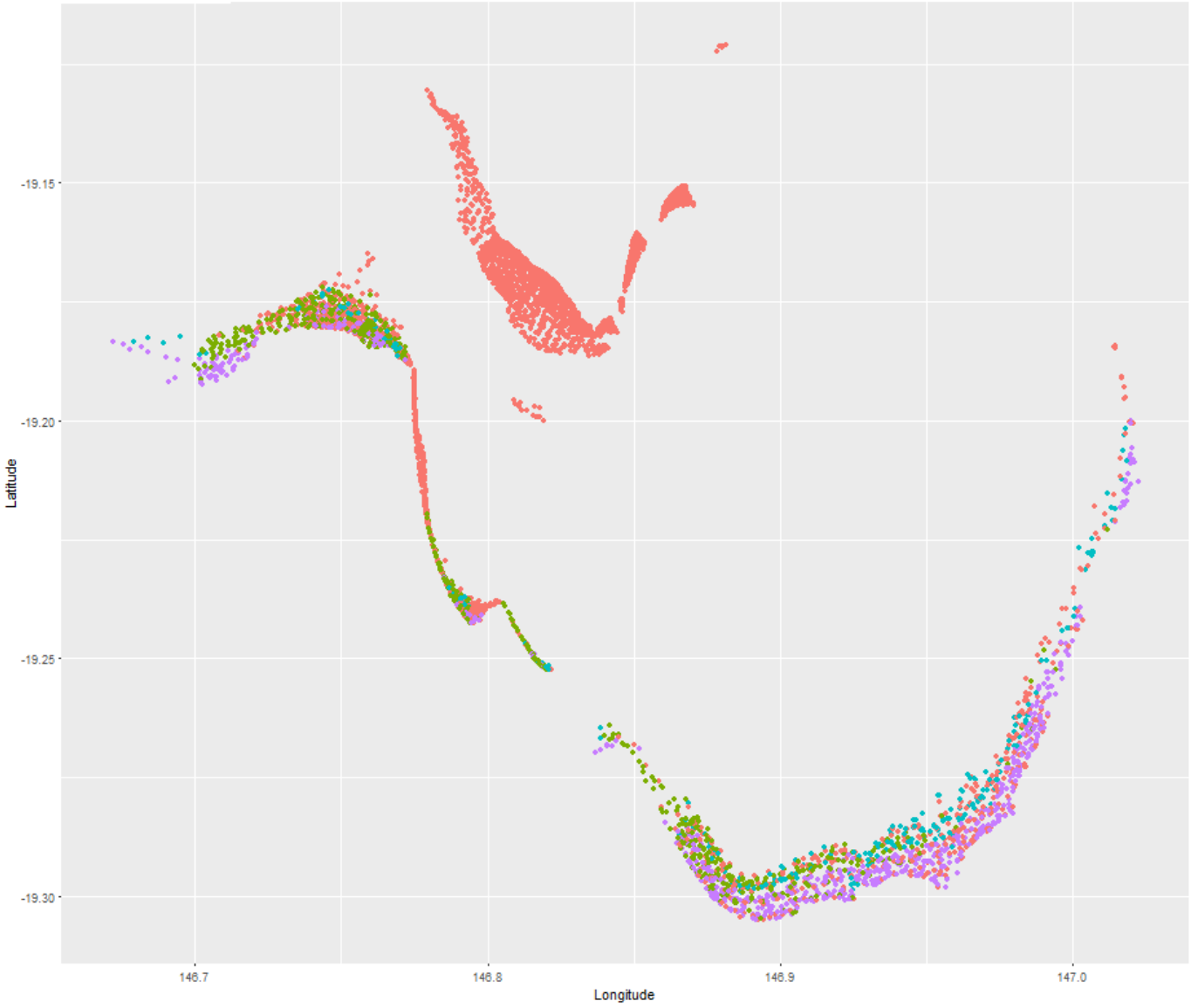

Figure 8: Node membership for all intertidal sites classified using the MRT on square-root transformed biomass. Colour code for nodes: ● = 2, ● = 4, ● = 6, ● = 7.

3b. MRT where the habitat is subtidal

Fitting an MRT to just the subtidal data results in only three terminal nodes based on depth and water type. The cross-validated relative error is much higher here (0.977) compared to the intertidal analysis (0.876). While there seems to be a small proportion of sites with higher *Cymodocea* and *Halodule* biomass, the MRT has found it difficult to distinguish sites overall, with the majority of sites having mostly a low biomass of all species.

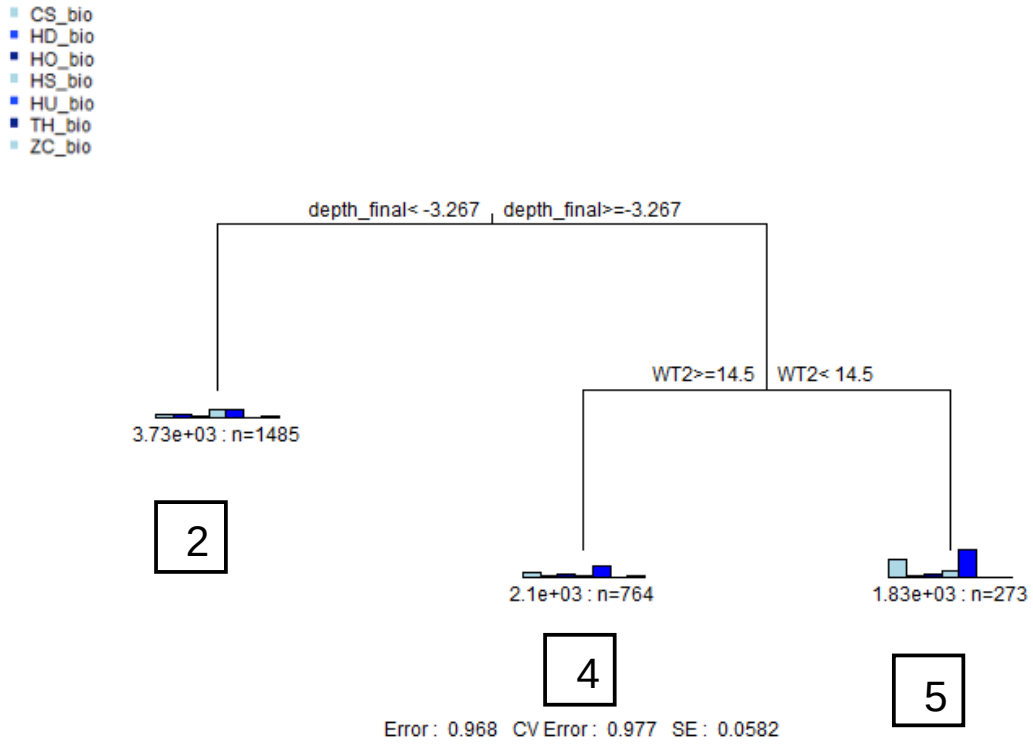

Figure 9: MRT on shallow and deep subtidal data where the response matrix is square-root transformed biomass. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 10.

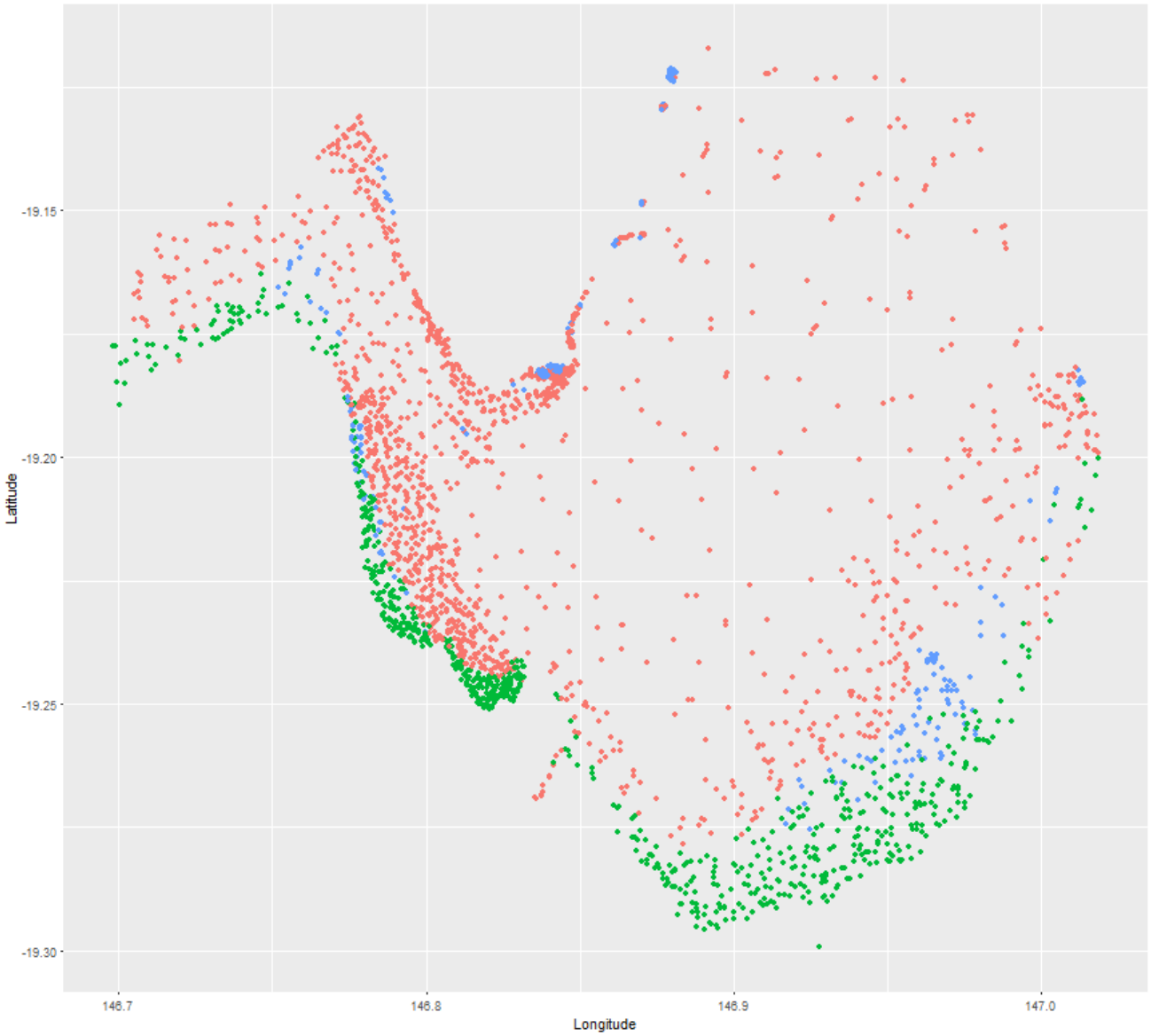

Figure 10: Spatial distribution of node membership for all shallow and deep subtidal sites classified using the MRT on square-root transformed biomass. Colour code for nodes: ● = 2, ● = 4, ● = 5.

4. Separate MRT for each year

We only show the maps of the clusters for the analyses of the individual years (not the trees) (Figure 11 and Figure 12).

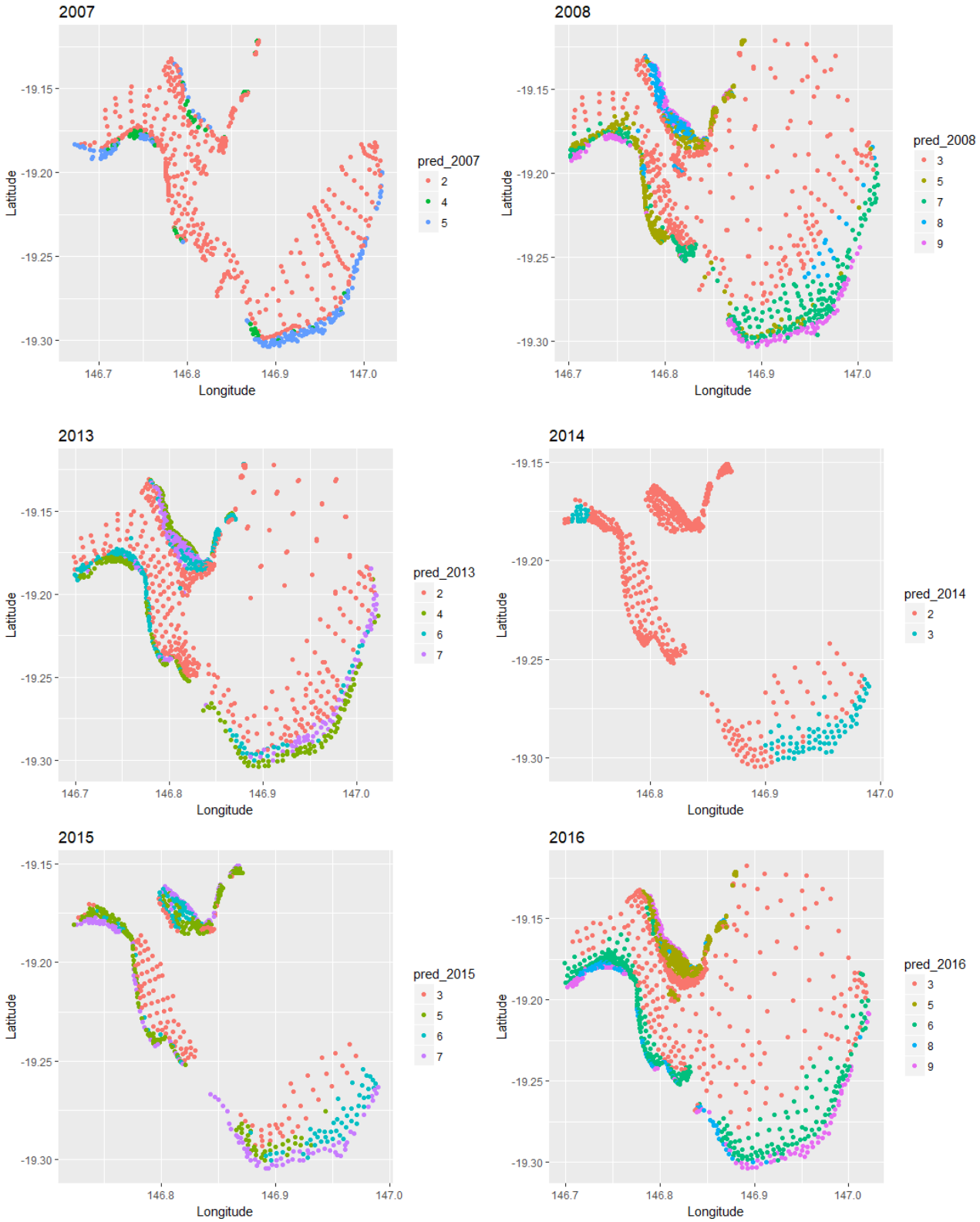

Figure 11: Spatial distribution of node membership for shallow and deep subtidal data where the response matrix is square-root transformed biomass every 'good year has been analysed separately. These would all be considered years of relatively "good" seagrass.

In the 'good' years, with the exception of 2014 (and to a lesser extent 2007 where there is less data), there are 4-5 nodes per tree and the distribution of communities is similar from map to map (Figure 11).

For the individual years the cluster numbering is not comparable from year to year i.e. Node 5 in 2007 will not be the same as Node 5 in 2008. However, Node 7 in 2015 is very similar to Node 9 in 2016 and it is the patterns we are interested in.

In the poor seagrass years (roughly 2009-2012), there is a lot less differentiation in community types with each year only having two nodes (Figure 12). The two nodes identified are very consistent from year to year.

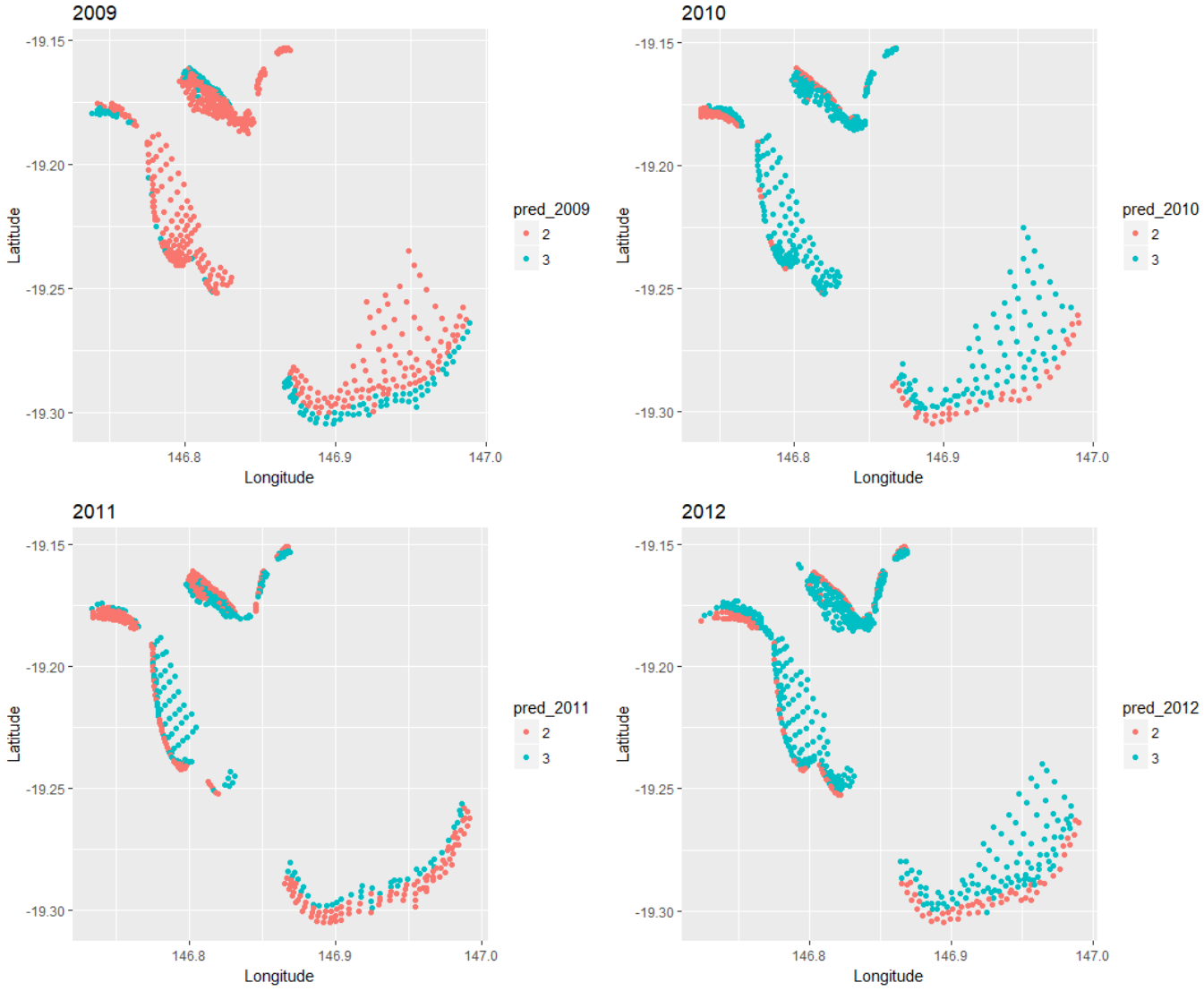

Figure 12: Spatial distribution of node membership for shallow and deep subtidal data where the response matrix is square-root transformed biomass and 2009, 2010, 2011 and 2012 have been analysed separately. These would all be considered years of relatively "poor" seagrass.

5. MRT for combined 'good years'

All habitats

Given we are interested in quantifying the desired state, which would by definition be a state based on years of high seagrass abundance, we have excluded the years 2009 to 2012 from the remainder of the tree analyses. Figure 13 shows the MRT on the data, excluding 2009-2012, where the response matrix is the square-root transformed matrix of biomass. Removing these years from the data has reduced the cross validated error from 0.894 to 0.842. While similar variables are being using to split the tree, there are now six nodes compared to the four in the Base-MRT. By removing some of the "noise" from the data, the tree is able to better quantify the relationships between species biomass' and the environmental predictors.

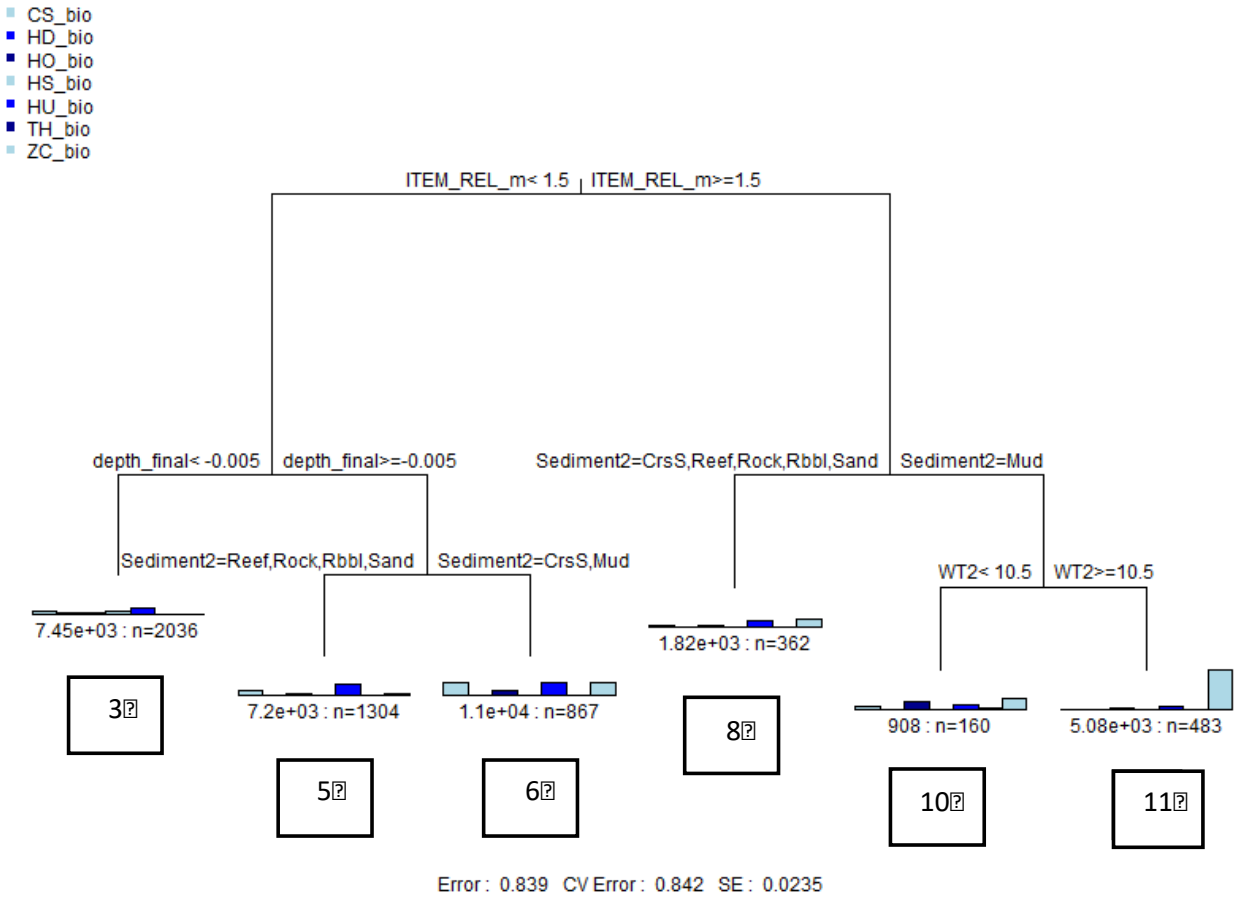

Figure 13: MRT on data excluding 2009-2012 where the response matrix is square-root transformed biomass separately. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 14.

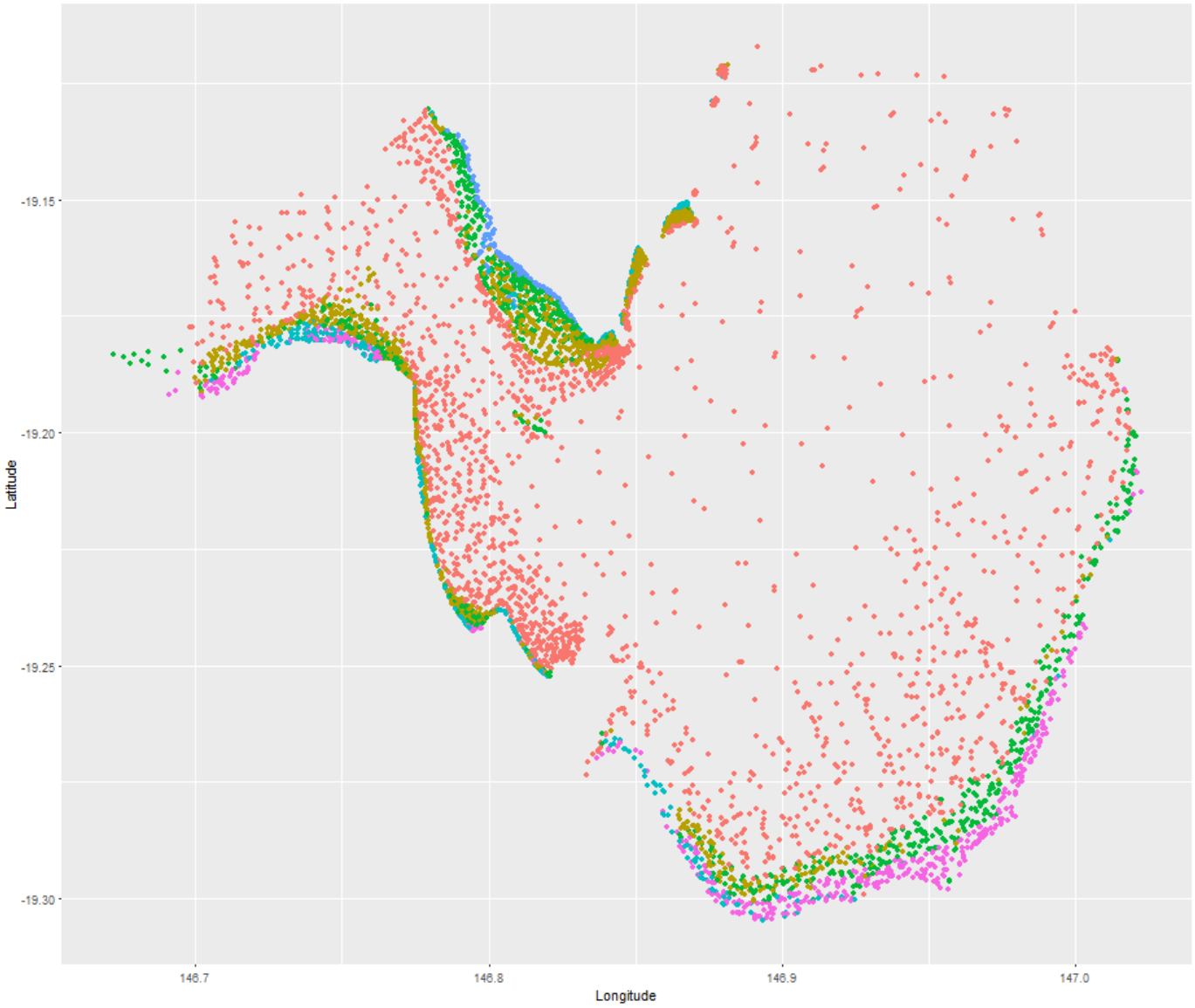

Figure 14: Spatial distribution of node membership for all sites excluding 2009-2012, classified using the MRT on square-root transformed biomass. Colour code for nodes: ● = 3, ● = 5, ● = 6, ● = 8, ● = 10, ● = 11.

Intertidal habitats

Similarly, repeating the intertidal analysis excluding the 2009-2012, the splits are similar but there is an additional node and the cross validated error has reduced by around 0.04 (Figure 15 and Figure 16).

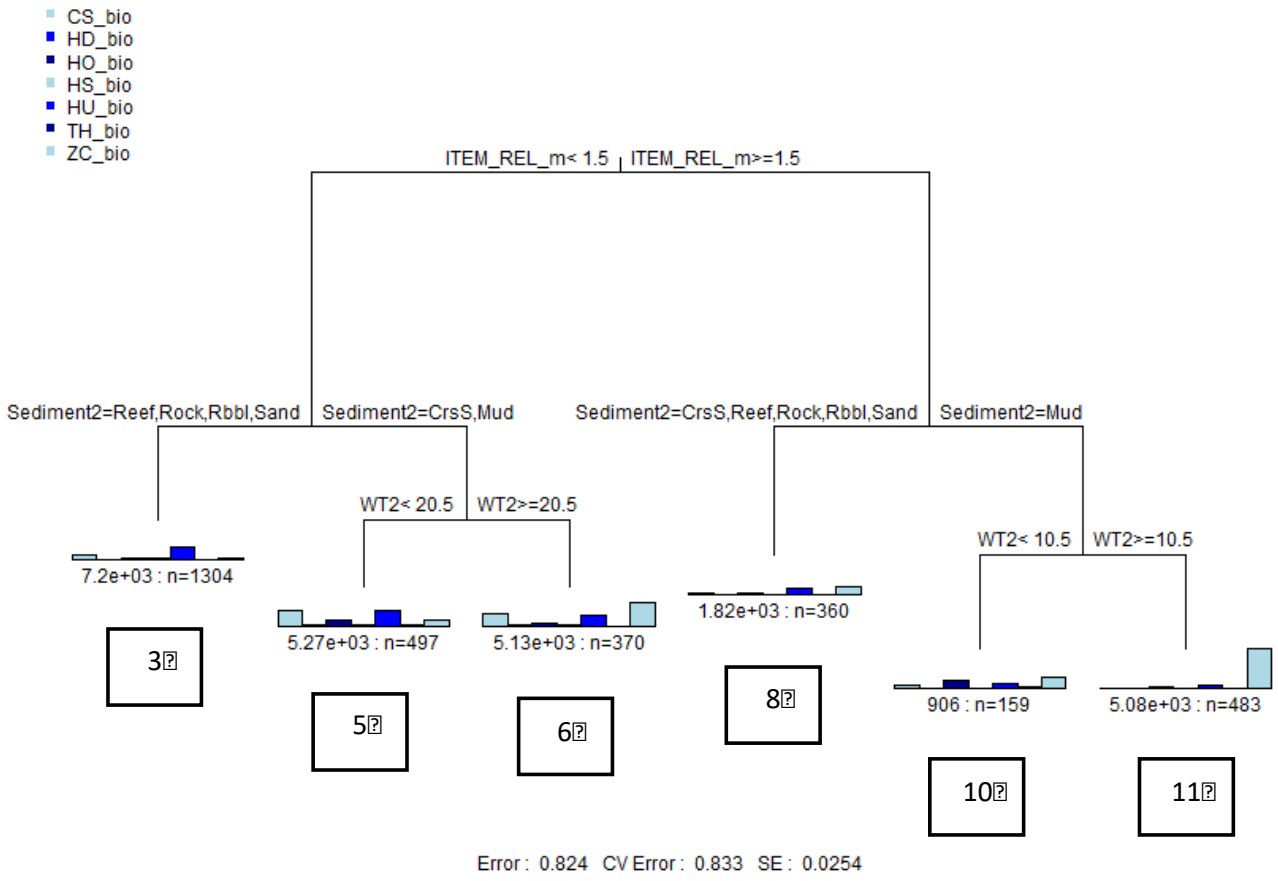

Figure 15: MRT on Intertidal data excluding 2009-2012 where the response matrix is square-root transformed biomass. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 16.

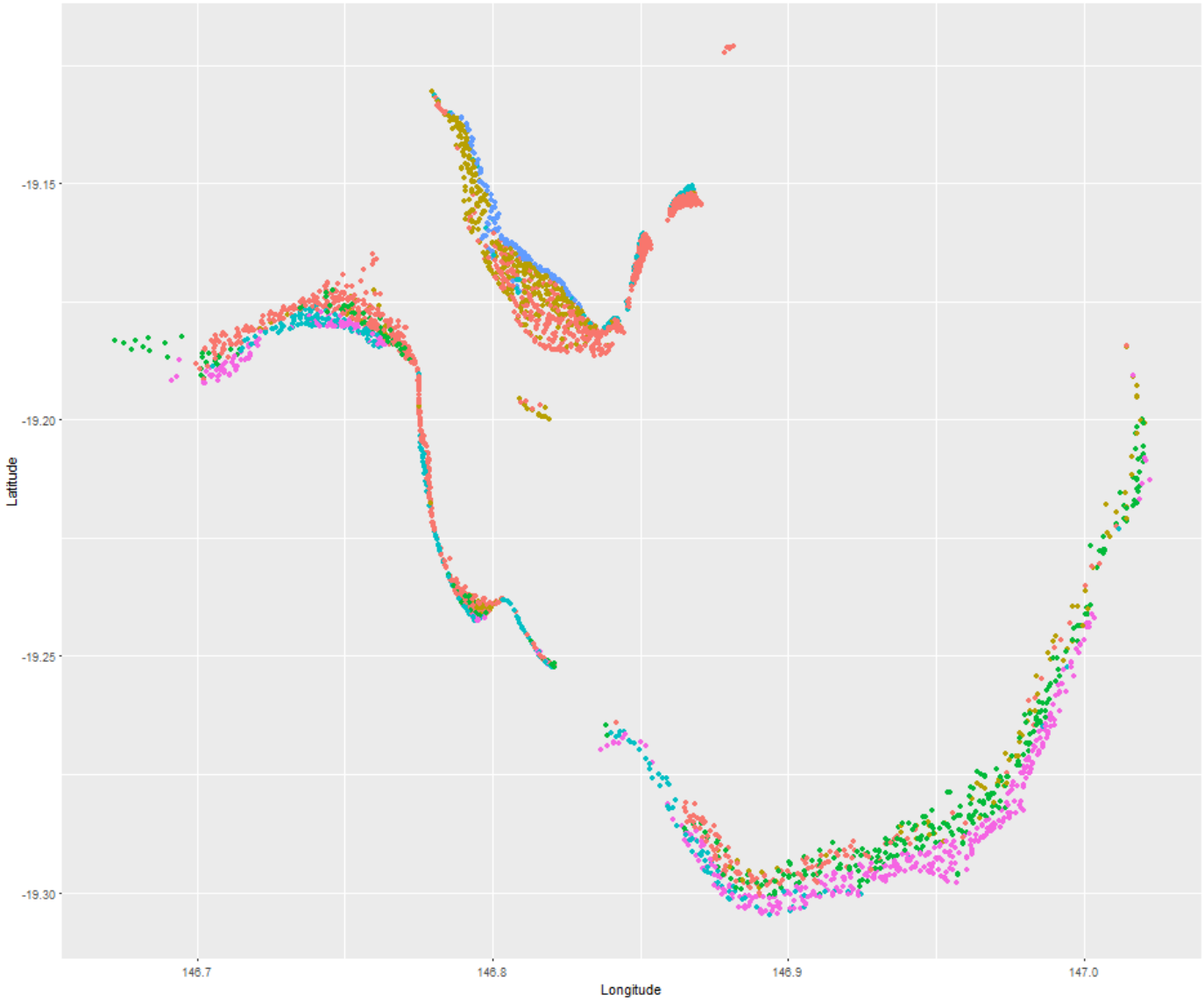

Figure 16: Spatial distribution of node membership for all intertidal sites excluding 2009-2012, classified using the MRT on square-root transformed biomass. Colour code for nodes: ● = 3, ● = 5, ● = 6, ● = 8, ● = 10, ● = 11.

Subtidal habitats

Repeating the subtidal analysis removing the 2009-2012 data, the first split on depth is the same but this time there is no further splitting of the tree (Figure 17 and Figure 18). The cross validated error is still very high.

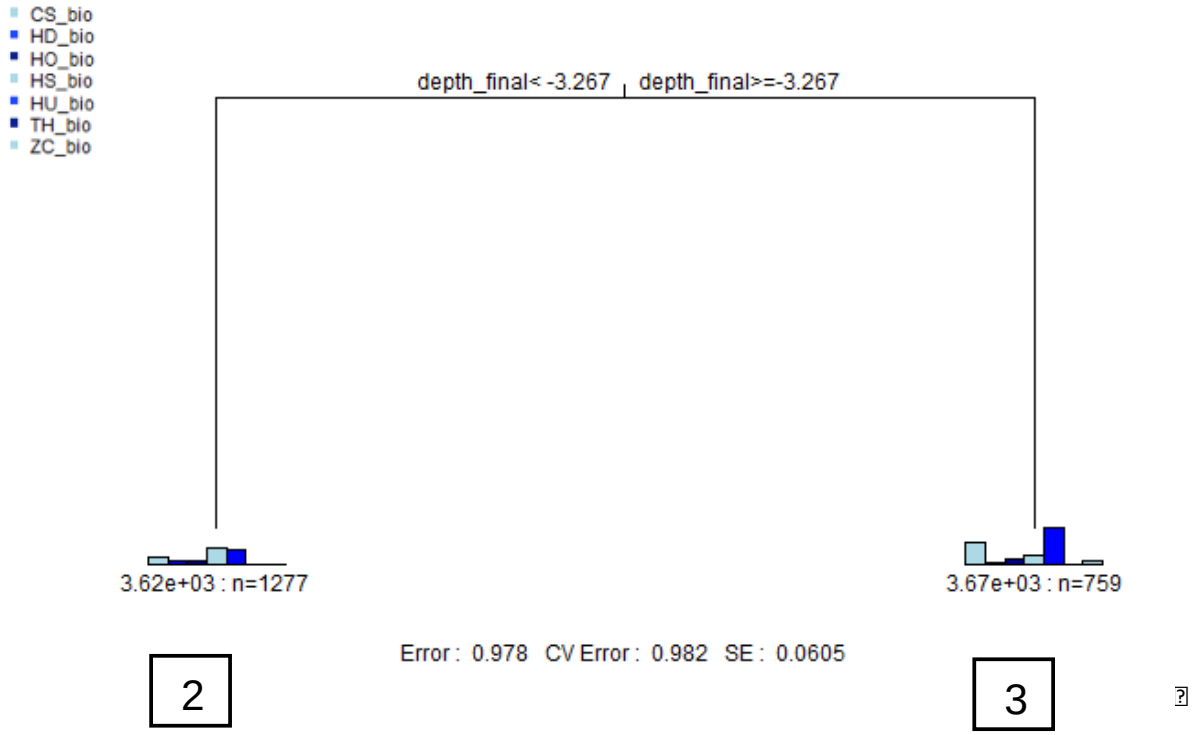

Figure 17: MRT on Subtidal data excluding 2009-2012 where the response matrix is square-root transformed biomass. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 18.

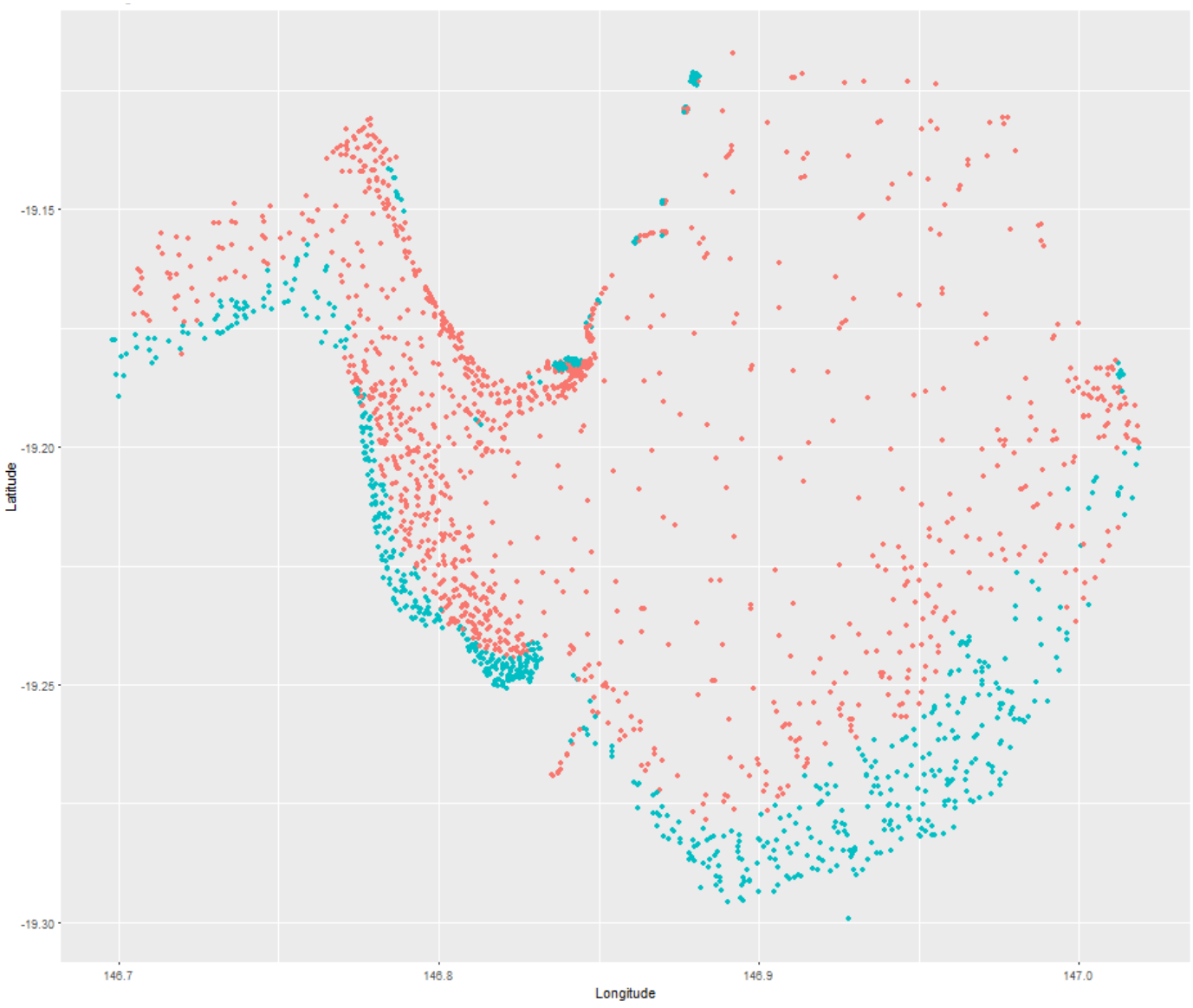

Figure 18: Spatial distribution of node membership for all subtidal sites excluding 2009-2012, classified using the MRT on square-root transformed biomass. Colour code for nodes: ● = 2, ● = 3.

6. MRT for combined 'good years' based on presence-absence

All habitats

We then repeated the previous three tree analyses with the presence-absence matrix as the response. Using the entire dataset, the splits in the tree (Figure 19) are almost the same as when the transformed biomass matrix was the response (Figure 13). The only exception is a different depth split. The intertidal analysis is again, very similar with an additional node (Figure 21 and Figure 22) and the subtidal analysis has a very minor difference in depth split (3.497m compared to the previous 3.267m; Figure 23 and Figure 24). The results appear to be quite robust to changes in the choice of response variable.

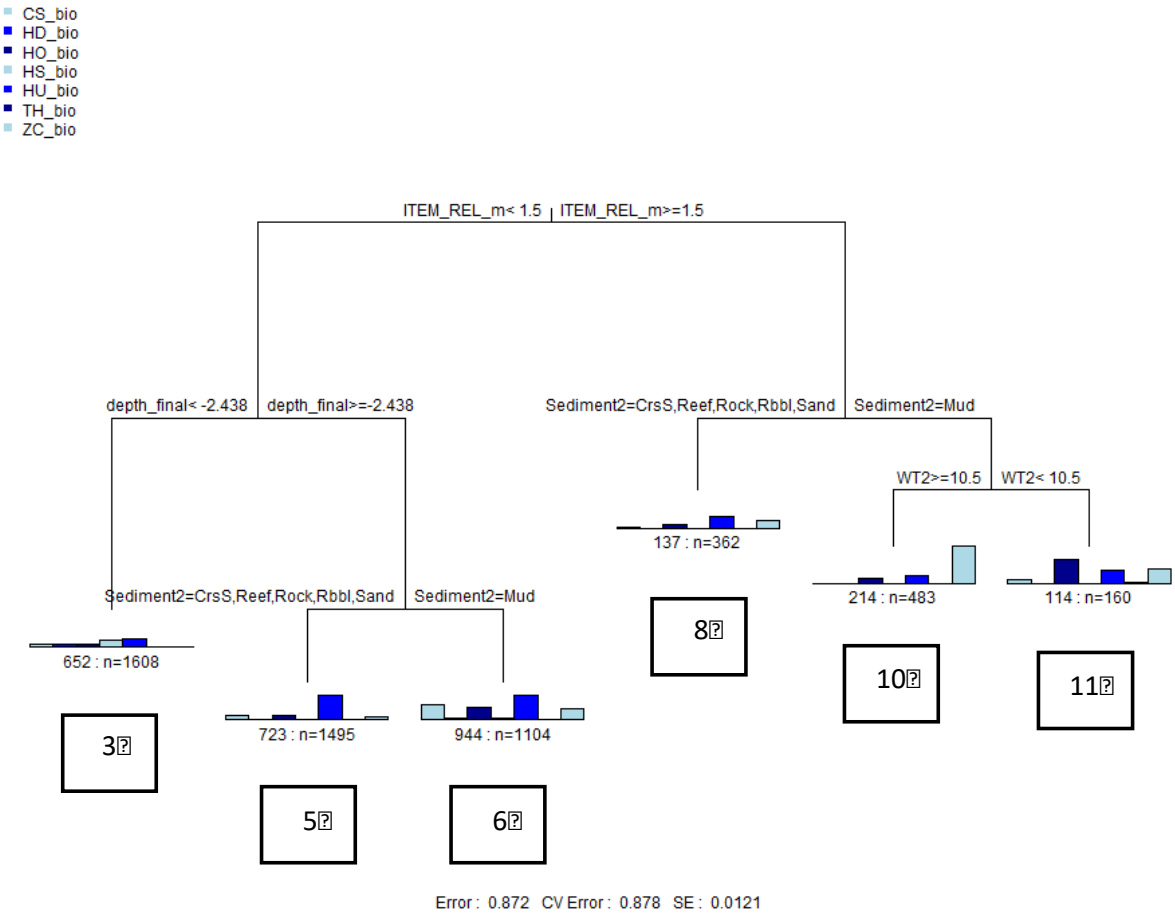

Figure 19: MRT on all data excluding 2009-2012 where the response matrix is presence-absence. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 20.

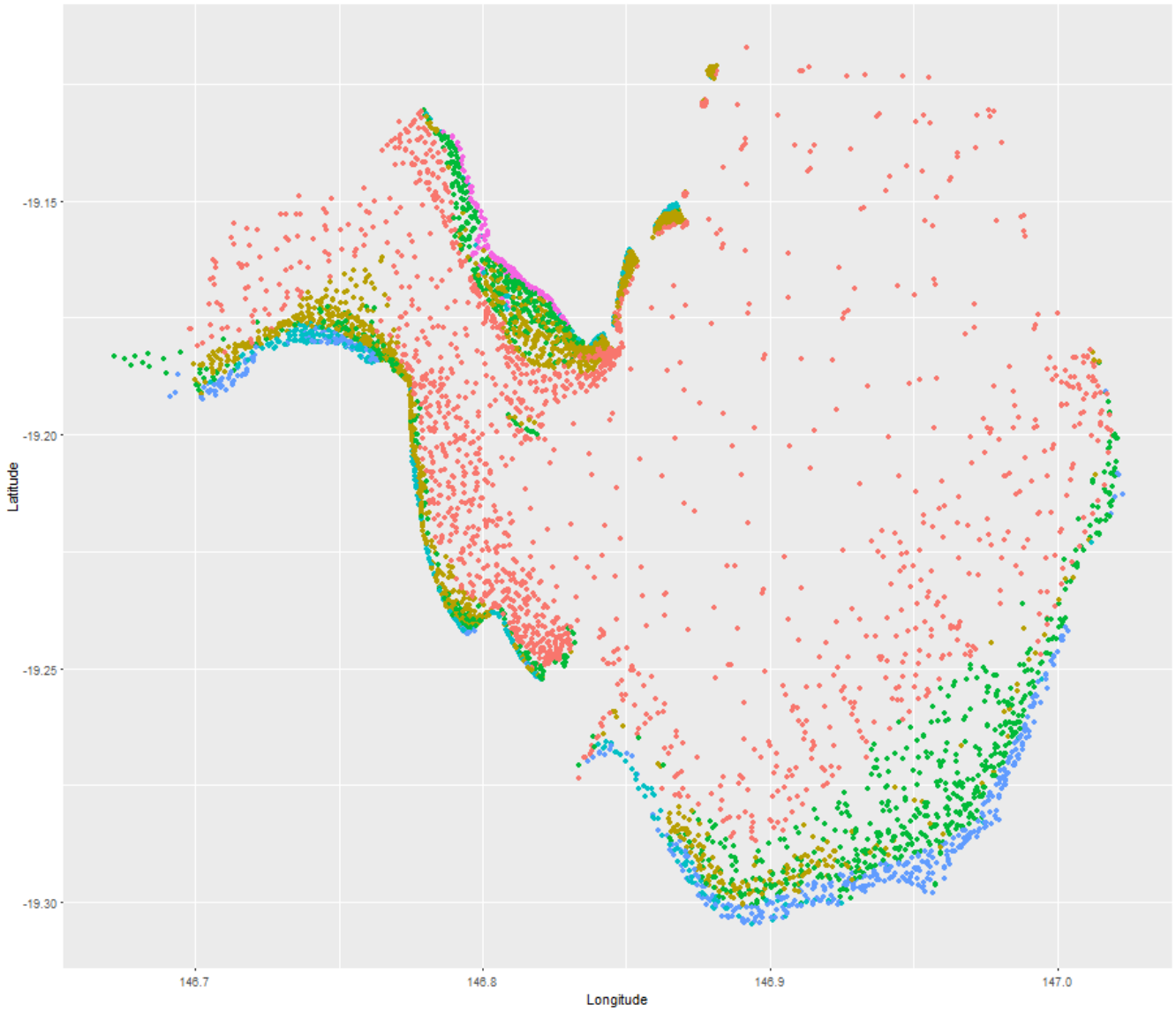

Figure 20: Spatial distribution of node membership for all sites excluding 2009-2012, classified using the MRT on presence-absence data. Colour code for nodes: ● = 3, ● = 5, ● = 6, ● = 8, ● = 10, ● = 11.

Intertidal habitats

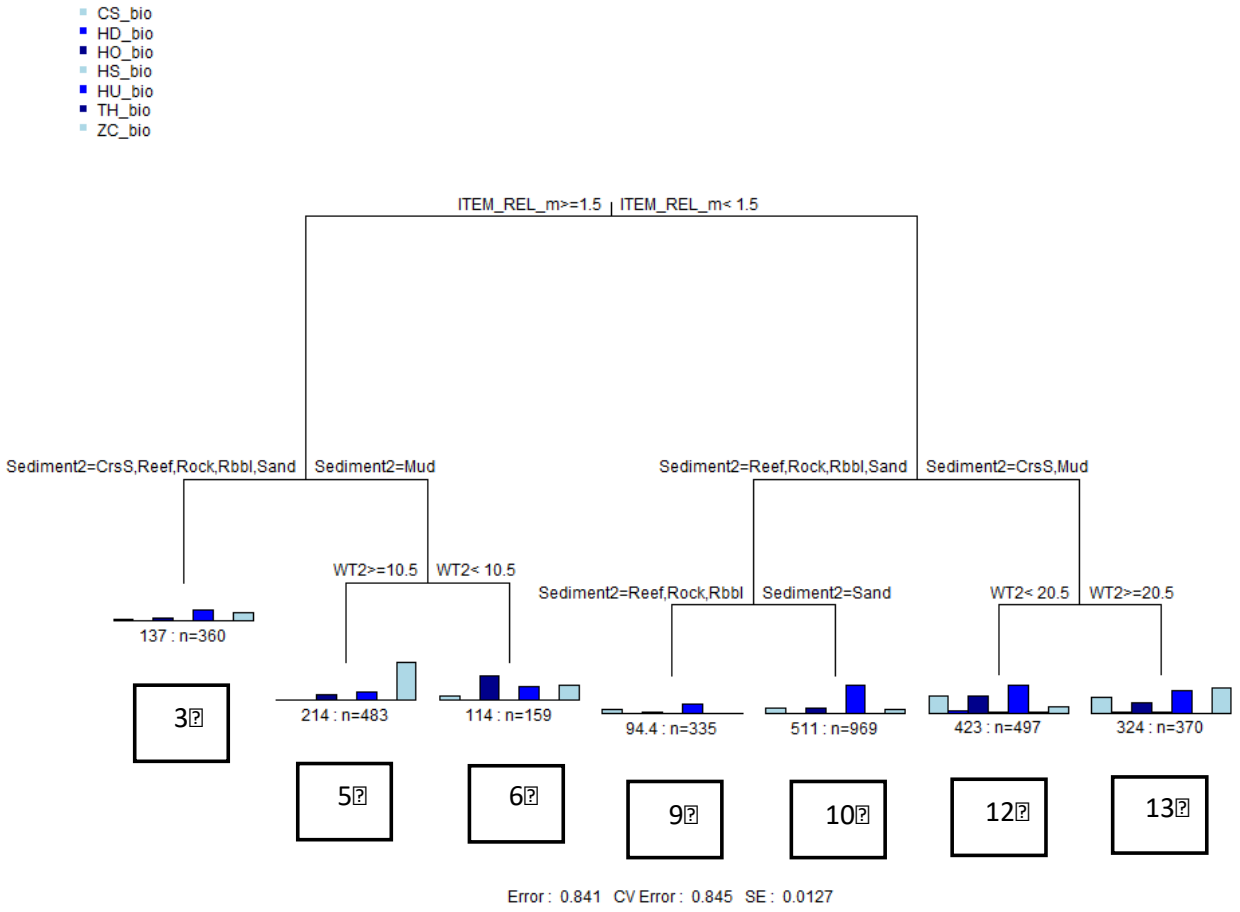

Figure 21: MRT on all Intertidal data excluding 2009-2012 where the response matrix is presence- absence. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 22. Figure 4.

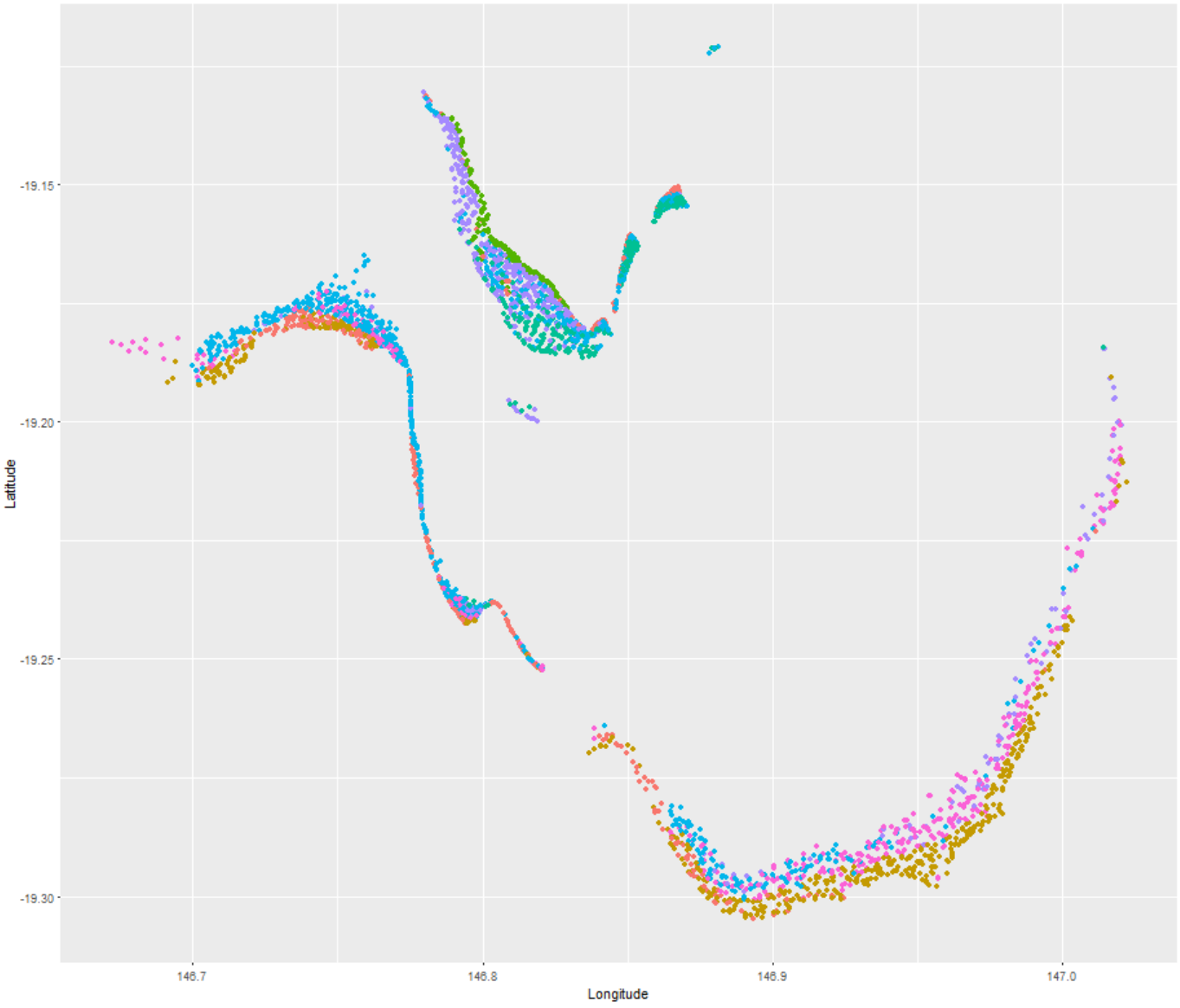

Figure 22: Spatial distribution of node membership for Intertidal sites excluding 2009-2012, classified using the MRT on presence-absence data. Colour code for nodes: ● = 3, ● = 5, ● = 6, ● = 9, ● = 10, ● = 12, ● = 13.

Subtidal habitats

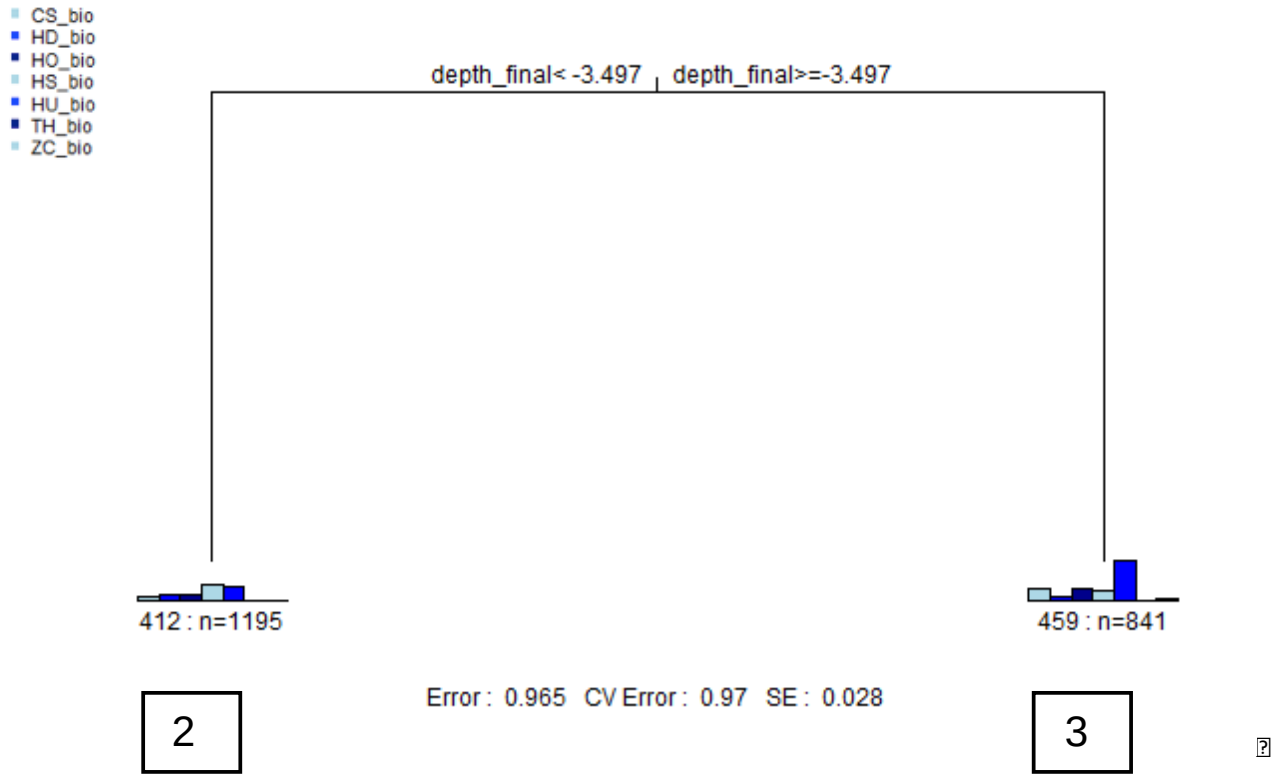

Figure 23: MRT on Subtidal data excluding 2009-2012 where the response matrix is presence-absence. The branches describe the variables used to split the tree. The numbers in the boxes are the labels for the terminal nodes and correspond to the labels in Figure 24.

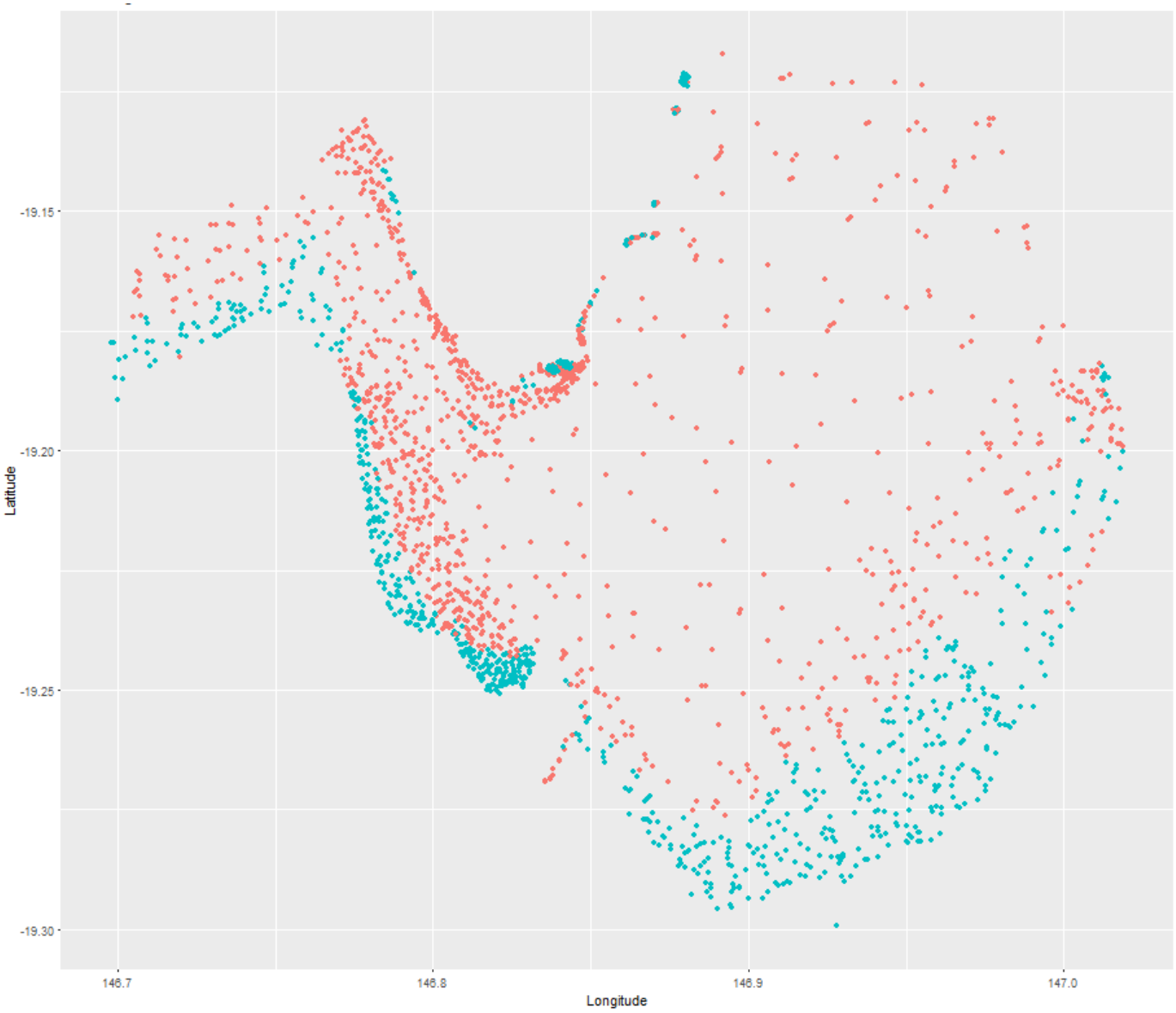

Figure 24: Spatial disribution of node membership for Subtidal sites excluding 2009-2012, classified using the MRT on presence-absence data. Colour code for nodes: ● = 2, ● = 3.

7. Community composition in 'good years'

The remainder of the analysis to estimate the abundance of each community type is based on the nodes determined by the MRTs on the presence-absence Intertidal and Subtidal data excluding 2009-2012 (Figure 21 and Figure 23). Further discussion about why this is the case can be found in the Discussion section.

In Figure 25 - Figure 27 we show the distribution of biomass values recorded for each species in each of those nodes. The Figures are box and whisker plots so the centre solid line is the median and the dots are considered to be outliers. Nodes 3 and 5 are very similar, with the main difference being the domination of *Zostera* in Node 5. Node 3 and 6 are also similar with a higher upper quartile (top of the box) for *Halophila ovalis, Halodule and Zostera*. Node 5 and 9 are very similar with a few more high records of *Cymodocea* in Node 9 and almost an absence of *Zostera*.

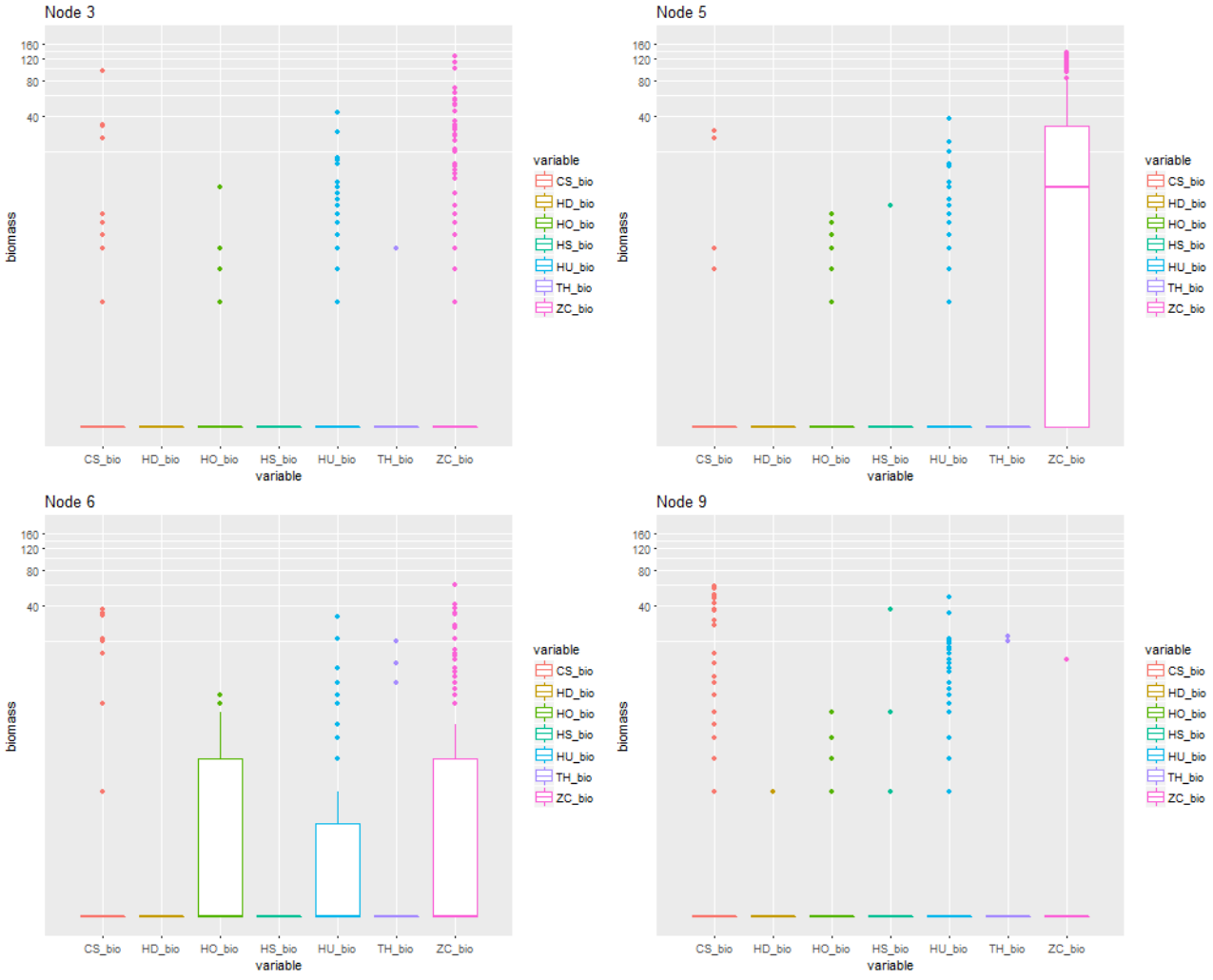

Figure 25: Box and whisker plot of the biomass observations recorded for each species in Node 3, 5, 6 and 9 for the MRT on intertidal data, excluding 2009-2012, where the response matrix is presence-absence. The y axis is on a log-scale (0.1 was added to all biomass values to accommodate 0 values).

Cymodocea and *Halodule* start to dominate more in Nodes 10 – 13 (Figure 26). Nodes 10 and 12 are very similar, except for a higher abundance of *Cymodocea* and *Halophila Ovalis* in Node 12. Node 10 and 13 are also similar except for less *Halodule* and a lot more *Zostera* and *Cymodocea* in Node 13.

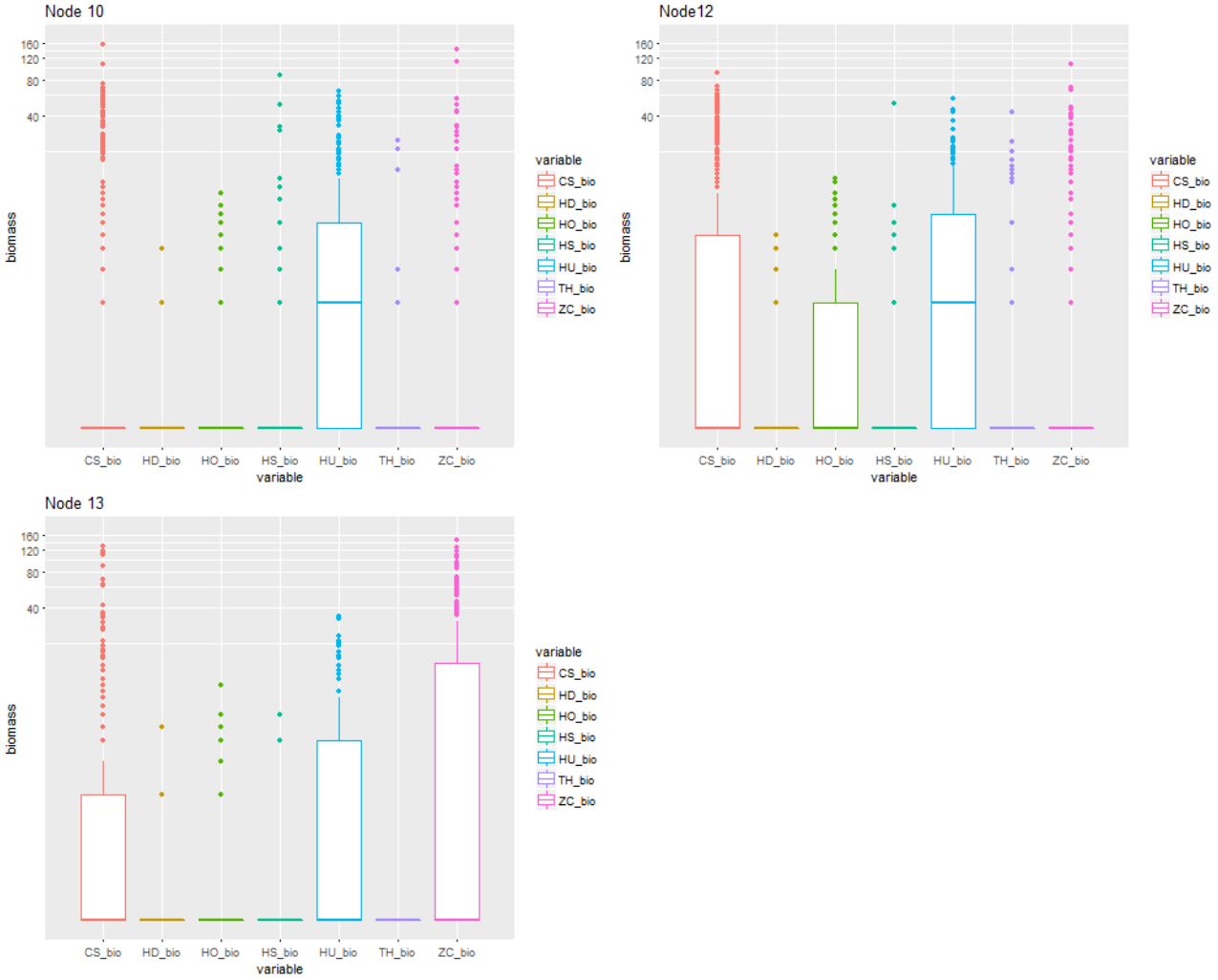

Figure 26: Box and whisker plot of the biomass observations recorded for each species in Nodes 10, 11 and 12 for the MRT on intertidal data, excluding 2009-2012, where the response matrix is presence- absence. The y axis is on a log-scale (0.1 was added to all biomass values to accommodate 0 values).

In the subtidal region, the two nodes are very similar with a few higher biomass values for *Halodule* and *Cymodocea Serrulata in Node 3*. The environmental gradient describing the distinction between Node 2 and Node 3 is the water depth (deeper water in Node 2).

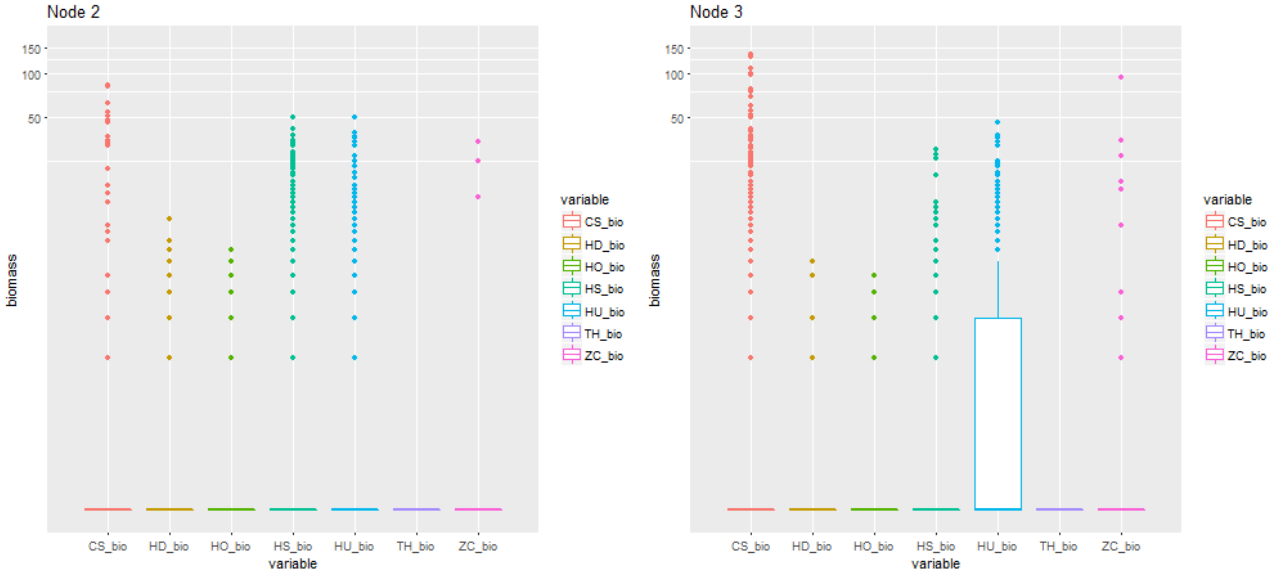

Figure 27: Box and whisker plot of the biomass observations recorded for each species in each node for the MRT on subtidal data, excluding 2009-2012, where the response matrix is presence- absence. The y axis is on a log-scale (0.1 was added to all biomass values to accommodate 0 values).

Part II: Temporal analysis

Methods

Once the species communities were established using the MRT method, we revisited the original dataset in entirety to determine the best method to arrive at a robust temporal trend in each cluster (tree node). While the final nodes were determined using the presence-absence data excluding 2009-2012, the trends are based on the total estimated biomass data for all species in a given node. To do this we had to use the relevant trees to predict the node membership of the observations from 2009-2012, based on the environmental covariates. The dataset was subset to observations collected from September through to December as this is when seagrass conditions would be expected to be at their best seasonally. We initially plotted the raw biomass values (total across all species) in each node by year to see how the nodes may differ in trend.

We then did an exploratory analysis on the trend in each node by fitting a series of simple Generalised Additive Models (GAMs; Wood 2017) to determine the best way of modeling the data. GAMs are a class of statistical models where the usual linear relationship between the response and predictors are replaced by several non-linear smooth functions to capture the non-linearities in the data. The GAMs were fitted using the mgcv package in R (Wood 2017).

The biomass data are “zero-inflated” meaning that a large proportion of the data are zero values. This typically makes model fitting more difficult than say normally distributed data as the data exhibit many zeros as well as a large number and range of positive values. Binomial models are a good way to model presence-absence data, however inference is limited to the likelihood of presence, providing no information about the mean biomass. Hurdle models are a class of models for count data that help handle excess zeros and over dispersion (Mullahy 1986). They are two-component models: a truncated count/continuous component, such as Poisson, gamma or negative binomial, for the positive values, and a hurdle component model for the zero vs. larger counts (binomial model). The results of the two models are combined for inference. Hurdle models allow for the environmental processes to be different for the two model components (modelling whether is seagrass is present vs how much is present).

A simpler alternative is to fit a model based on the Tweedie distribution (Tweedie 1984). Tweedie distributions are a family of probability distributions which include the continuous normal and gamma distributions, the discrete scaled Poisson distribution as well as the class of mixed compound Poisson-gamma distributions which have a positive mass at zero but are otherwise continuous. It is these mixture distributions which should provide a good representation of our data. For the Tweedie distribution, the user sets the ‘p’ value which indicates the degree of ‘mixing’.

To determine the best method for fitting the models, a series of GAMs were explored and the standard model diagnostics checked. The simple models considered were:

1. Binomial GAM (with logit link) to the presence-absence data
2. Gamma GAM (with log link) to the biomass data (given at least some present)
3. Tweedie GAM (with p=1.2) to all data at once. Note: other values of p were tested.

We then attempted to incorporate some environmental variables into our analysis to help explain the temporal trends. The variables that were used in the tree proved to be very

uninformative in the models and caused lots of model fitting problems as they had already been used to cluster the data. Unfortunately, the climate-data explanatory variables had the same value for every observation for every year. This means that the model cannot handle having both a year term and the other variables as the effects can't be differentiated. As we are trying to quantify the temporal trend it was necessary to leave in the year variable and so we were forced to drop all other variables from the analysis. This is equivalent to averaging over the values of the different covariates to determine the trend in the mean across time. The assumption here is that we are getting a "representative" sample of the environmental conditions in each year i.e. this approach does not account for if one year we sample all muddy sites and the next year we sample all sandy sites. This is highly unlikely to be a problem given we have already broken down the data into the different community types based on the environmental variables.

While GAMs are good for visualizing a trend and incorporating non-linear relationships between a predictor and the response, Generalized Linear Models (GLMs) are a simpler modeling technique that can have better properties for inference. Given we were only left with 'Year' as a response variable (choosing to average all values collected in a given node for a given year (Sept-Dec)) we were able to instead fit GLMs. We did not consider incorporating a monthly component into our models as this would rely on the sampling being consistent across the months each year.

We fitted the GLMs for the two modeling approaches (1) hurdle and (2) Tweedie, noting that the hurdle would most likely provide a better fit but the Tweedie a simpler single model approach. While the estimate of the mean (by definition) makes very little difference between the two scenarios, the uncertainty estimates can be quite different. While preference should be for the simpler approach, if adequate, we calculated the uncertainty under both modelling approaches to ensure that the Tweedie is an adequate representation of the temporal uncertainty.

Under the Tweedie model the uncertainty was estimated by calculating the 95% confidence interval of model predictions for each year in a given region. For the hurdle model, the uncertainty around the index was calculated using a parametric bootstrap based on 5000 samples. For each bootstrap sample, each model stage (gamma and binomial) was fitted and predictions calculated for each 'Year'. The two sets of simulated predictions were multiplied together for each bootstrap sample to give the predicted catch rates. A 95% confidence interval was then calculated for each 'Year' by taking the 0.025% and the 0.975% percentiles from the bootstrap distribution for each 'Year'.

Results

We first show the results of the temporal trends in the nodes at the intertidal sites, followed by the subtidal.

1. Intertidal patterns

The intertidal raw data, broken down by node, show some general patterns (Figure 28). With the exception of Node 9, all of the nodes exhibit an obvious decline between 2006 and 2011 followed by a subsequent increase. The overall abundance of seagrass during the better years in Nodes 5, 10 and 13 is higher than the other nodes.

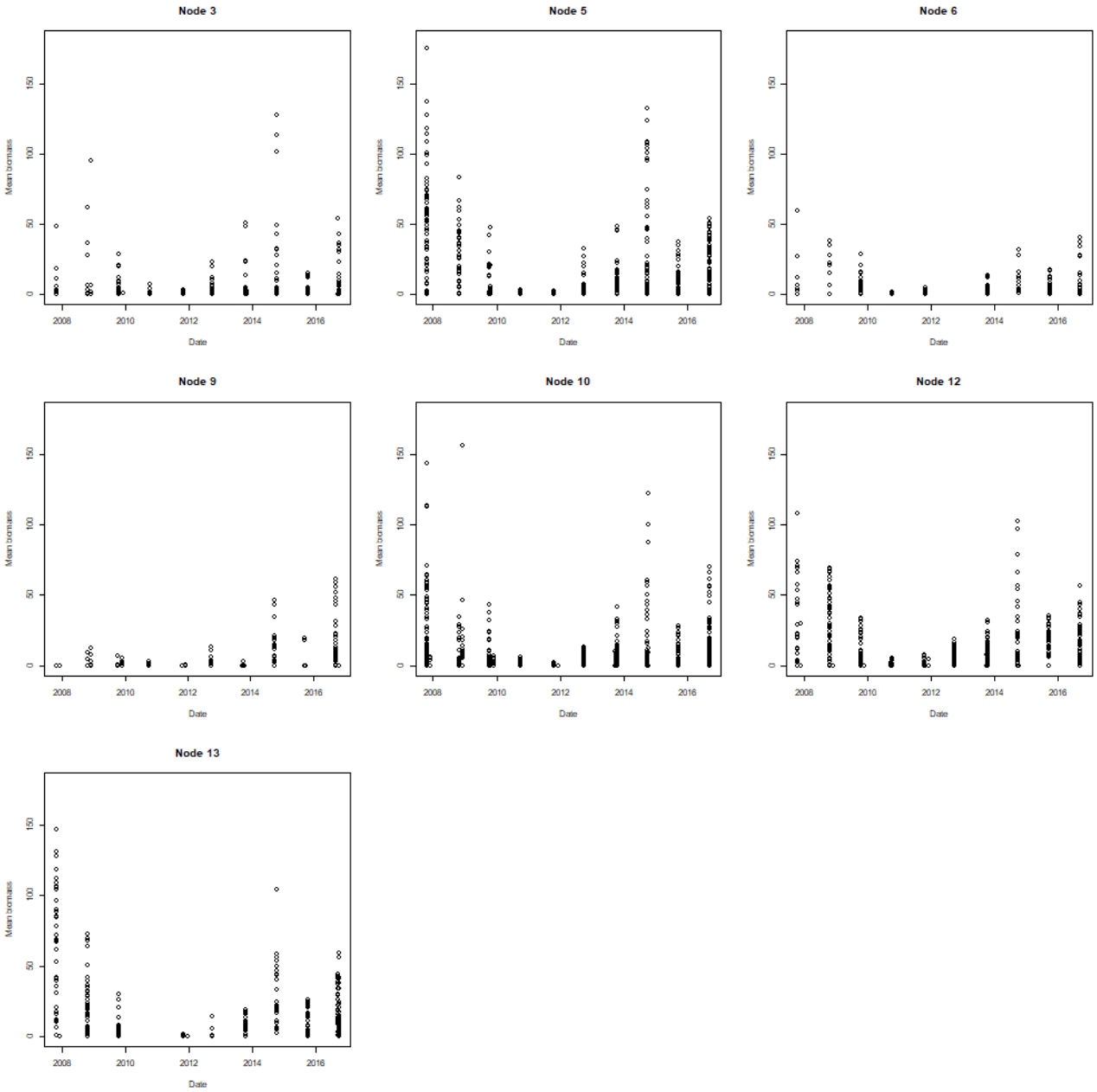

Figure 28: Plot of mean total biomass per site in each Node for the intertidal data

1a. Binomial GAM

The plot of the smooth term estimated by the simple binomial GAM fit to the presence-absence data in each node show a similar pattern to the raw data, with all except Node 9 showing a decline followed by an increase (Figure 29). As the response variable is presence-absence, the trends indicate that in general seagrass is observed at a decreasing proportion of sites between 2007 and 2011 and then the proportion later increases. While these smooth terms are all on the same scale and can be compared between Nodes, the absolute values of the y-axis can't easily be interpreted as the model was based on a logit transformation.

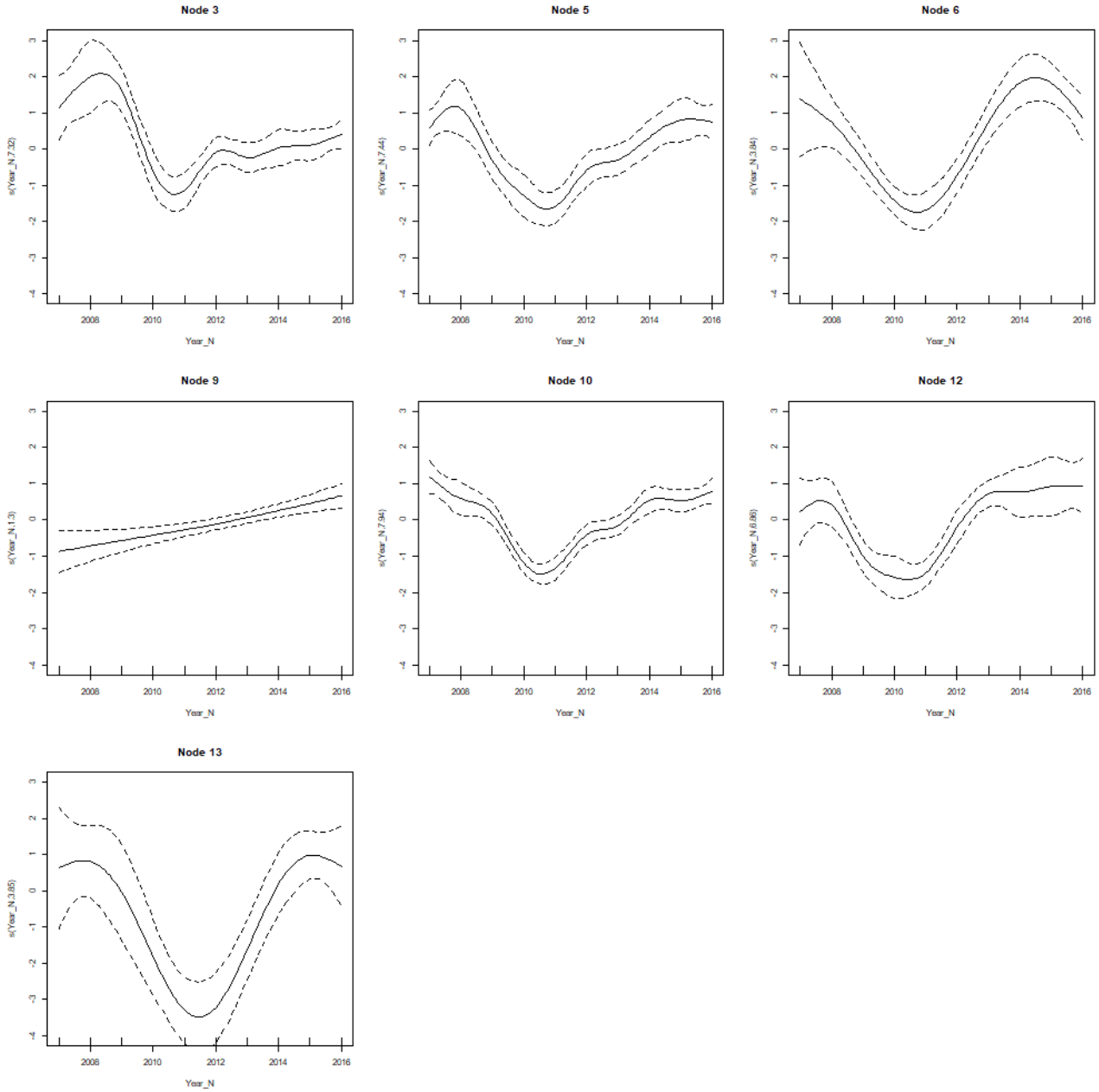

Figure 29: Plot of smooth term and 95% confidence intervals estimated by binomial GAM in each Node for the intertidal data

1b. Gamma GAM

The gamma GAMs provided a very good fit to the positive biomass data (biomass>0). Again, the general trends are of decreasing and then increasing abundances to varying extents in each of the nodes (Figure 30). Recall that the absolute value of the y-axis cannot be interpreted as the data have undergone transformation for the analysis, but the general trends are of interest. These plots show that not only do we observe seagrass at a decreasing then increasing number of sites, we also observe, at the sites where seagrass is present, a decreasing then increasing trend in the amount (total biomass) of seagrass. The confidence intervals around Node 3, 6 and 9 are larger than the other nodes, indicating that we have less confidence in these trends (due to less or more variable data in the node).

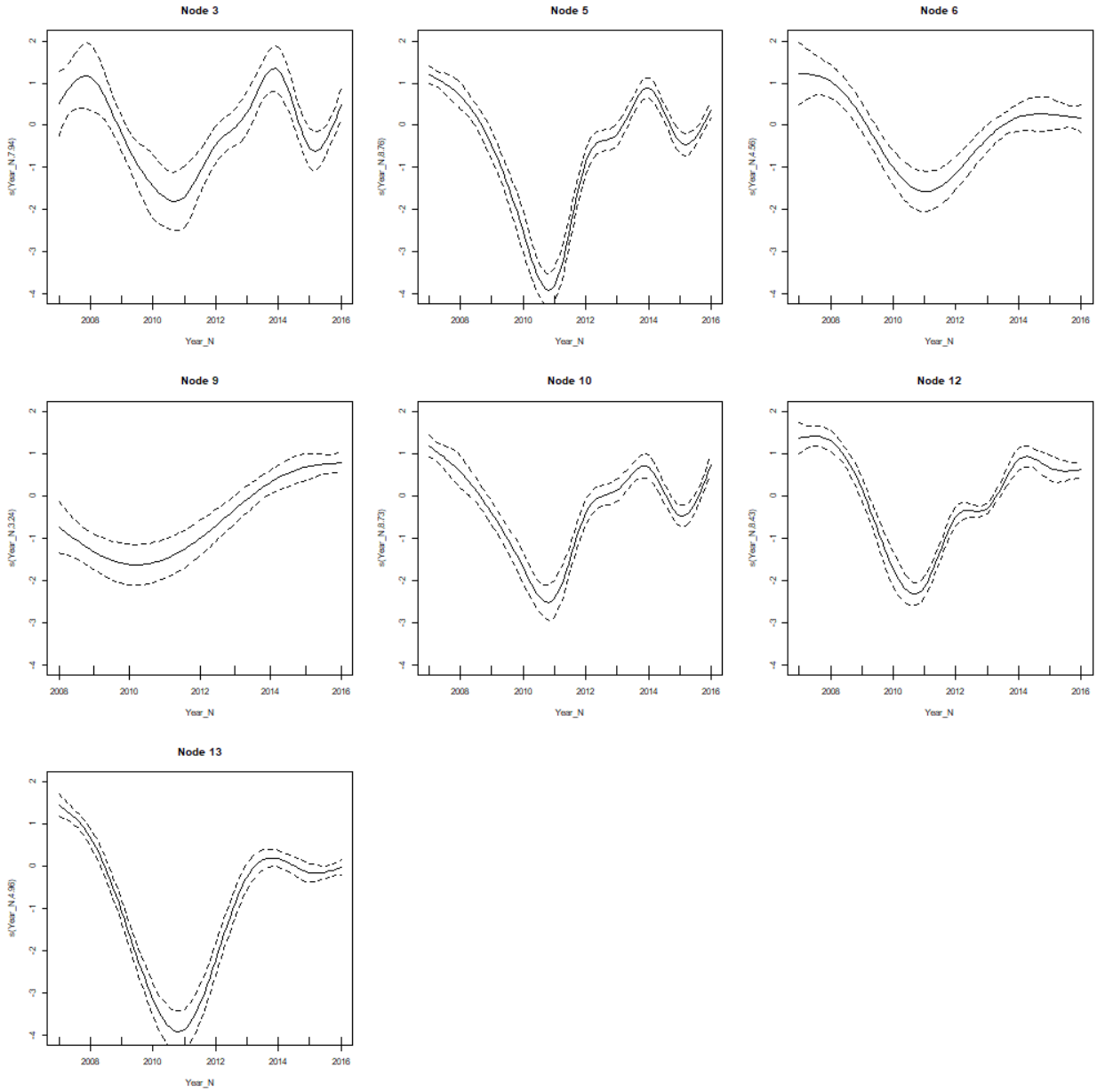

Figure 30: Smooth term fit estimated by simple Gamma model to positive biomass (>0) data in each node for intertidal data

1c. Tweedie model

The Tweedie models didn't fit the data as well as the gamma models, however this would be expected as the Tweedie models were fitting to both the zero and continuous values simultaneously. The model fits were adequate in all Nodes, based on standard model diagnostics. The trends are similar to the binomial and gamma models (Figure 31). The estimated smooth term for Node 9 shows the high level of uncertainty estimating this trend, in particular in the early and middle range of the trend the confidence intervals are very wide. Looking back at the raw data, this is because there is very little data in this Node during those points in time (Figure 28).

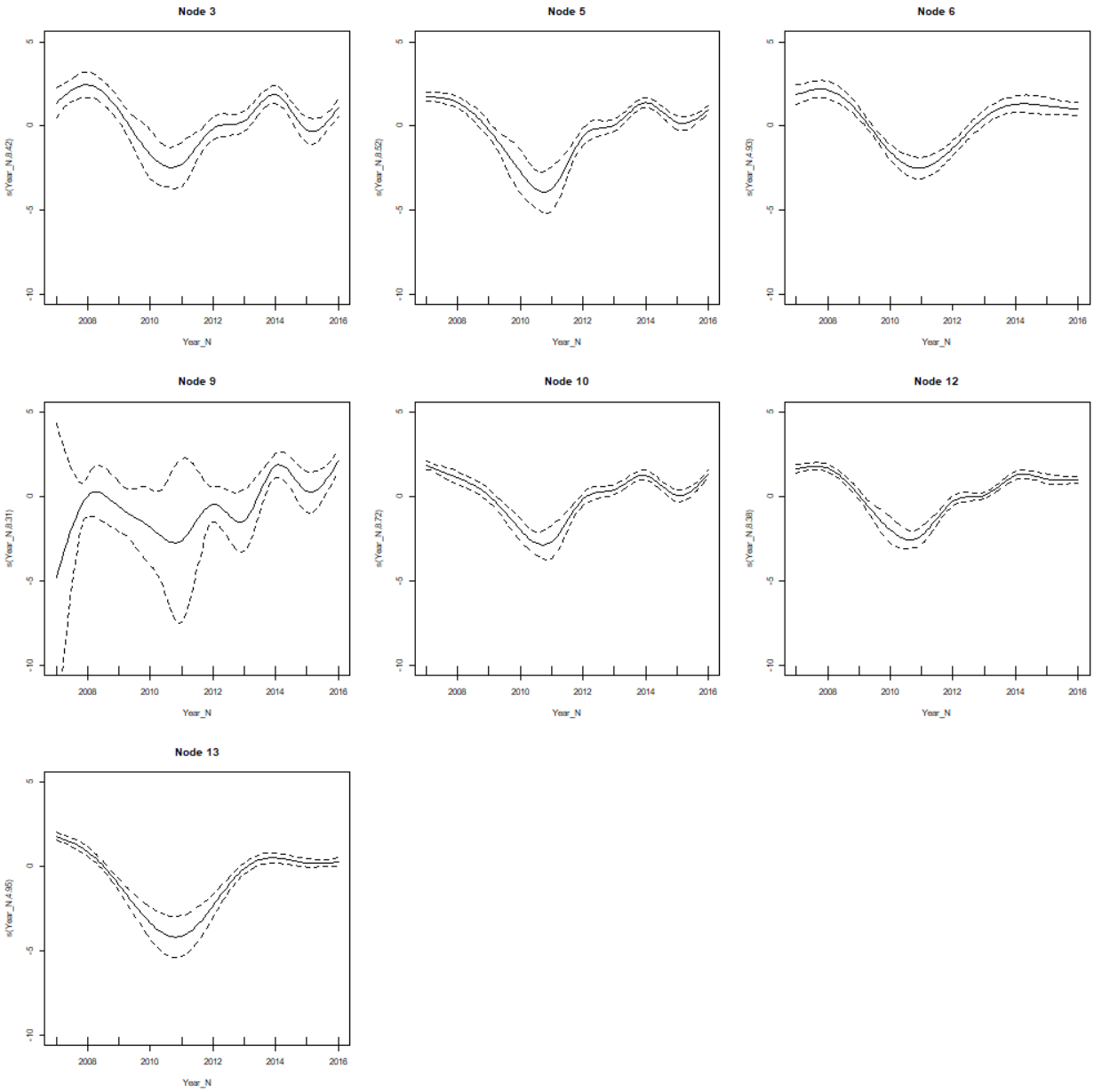

Figure 31: Smooth term fit estimated by Tweedie model to biomass data in each node for intertidal data

1d. Hurdle and Tweedie comparisons of estimated confidence

We then compared the mean and uncertainty under the two modelling approaches for Node 3, being a 'typical' node (Figure 32). The mean estimates under the approaches are almost identical and the uncertainty estimates similar but wider under the hurdle in some years (eg. 2008 and 2014) and wider under the Tweedie in other years (eg. 2007 and 2009). Given the Tweedie is not consistently the lower of the two (so we are not concerned it is an underestimate), we recommend and have proceeded using the Tweedie model to calculate the estimates and uncertainty for the remaining nodes. These predictions are on the response scales so the trends and estimates of uncertainty can now be interpreted in an absolute sense.

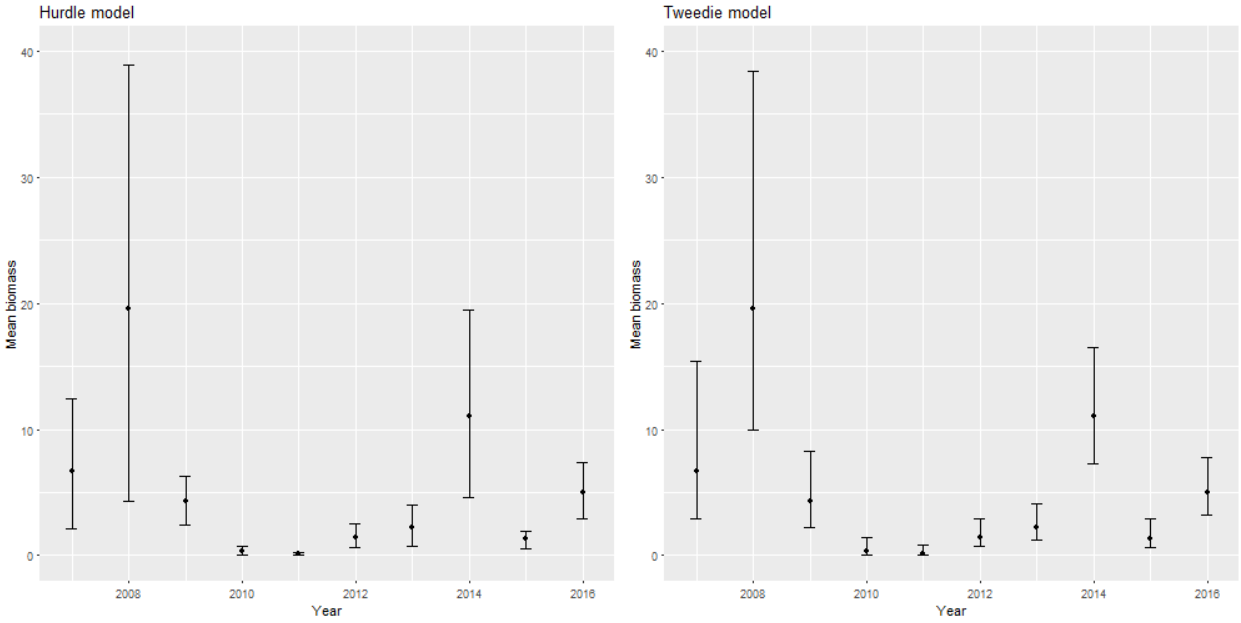

Figure 32: Comparison of the estimated confidence intervals under the hurdle and Tweedie modelling approaches in Node 3 for intertidal data

1e. Tweedie estimated trend and uncertainty for each node

The estimated trend and associated uncertainty in each of the nodes are shown in Figure 33. Where there is no uncertainty bound in the figure, there was insufficient data to estimate it (eg. Node 13 2007). While the magnitude of the 'peak' years varies from node to node, the poor years and trends are similar. Given the similarity in trends from node to node we expect that differences between the two estimation methods would be fairly consistent in the remaining nodes.

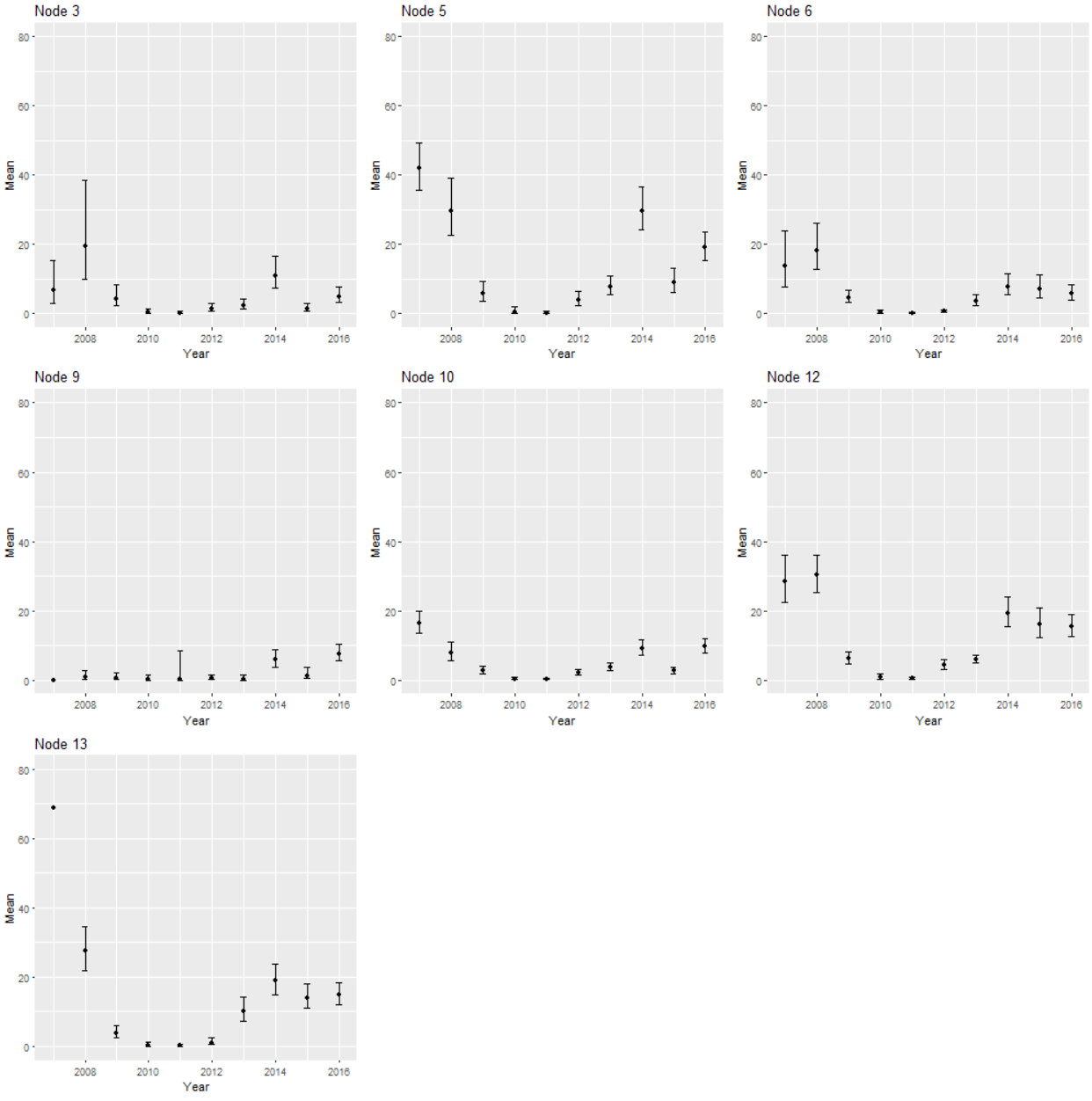

Figure 33: Estimated mean and 95% confidence intervals for mean biomass using the Tweedie model in the intertidal area

2. Subtidal patterns

In the subtidal area the raw mean biomass in the two nodes are similar with the most noticeable difference being a lower initial year (Figure 34).

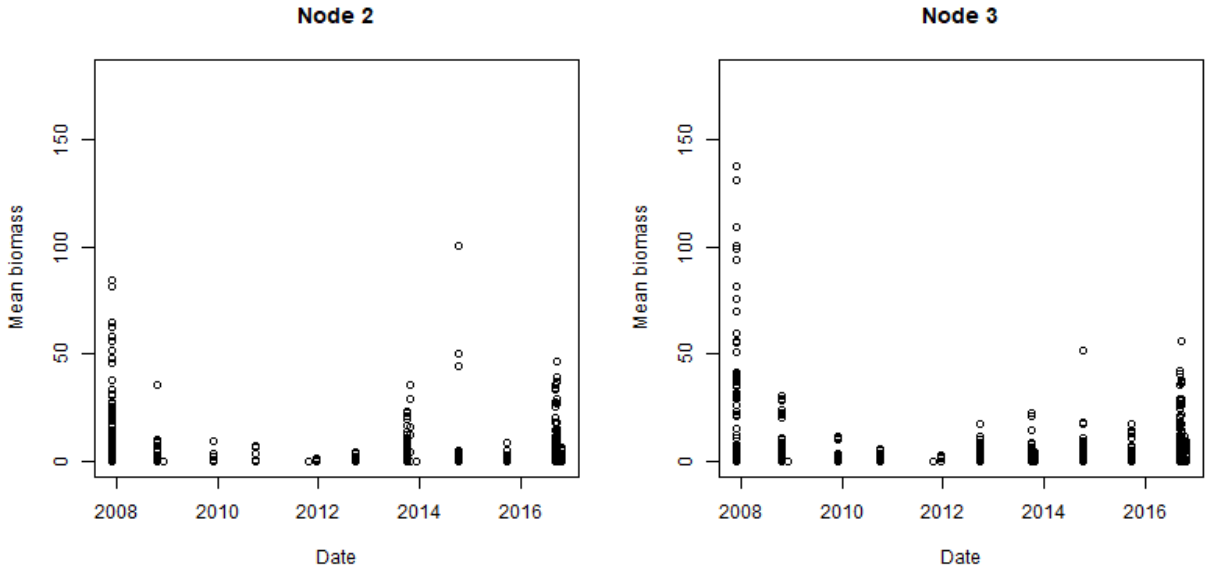

Figure 34: Plot of raw mean biomass data in the two nodes in the subtidal data

2a. Binomial GAM

The plot of the smooth term estimated by the simple binomial GAM fit to the presence-absence data in Node 2 (Figure 35) shows a similar trend to the intertidal sites. The trend indicates that seagrass is seen at a decreasing proportion of sites between 2007 and 2011 and then the proportion later increases. The trend is similar for Node 3 but the uncertainty around the estimates towards the centre dominate the figure as there is not much data to estimate the mean here. Recall that the absolute values of the y-axis can't easily be interpreted as the model was based on a logit transformation.

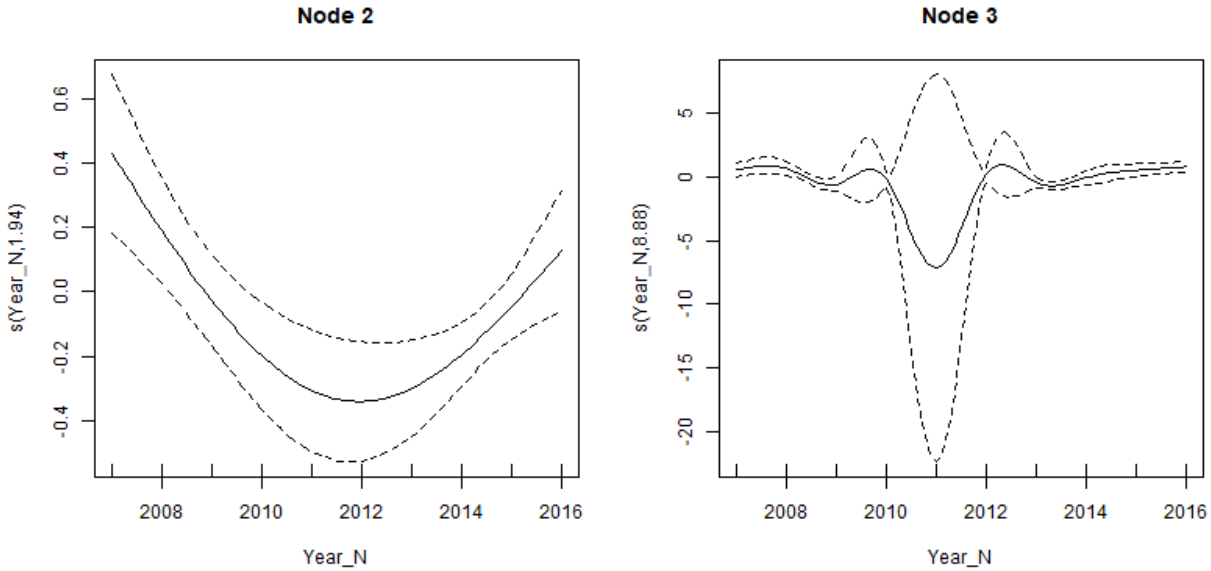

Figure 35: Plot of smooth term estimated by binomial GAM in each Node for subtidal data. Note the two figures have a different y-scale.

2b. Gamma GAM

The gamma GAMs also show the now consistent trends of decreasing and then increasing abundances, this time decreasing faster than they increase (Figure 36).

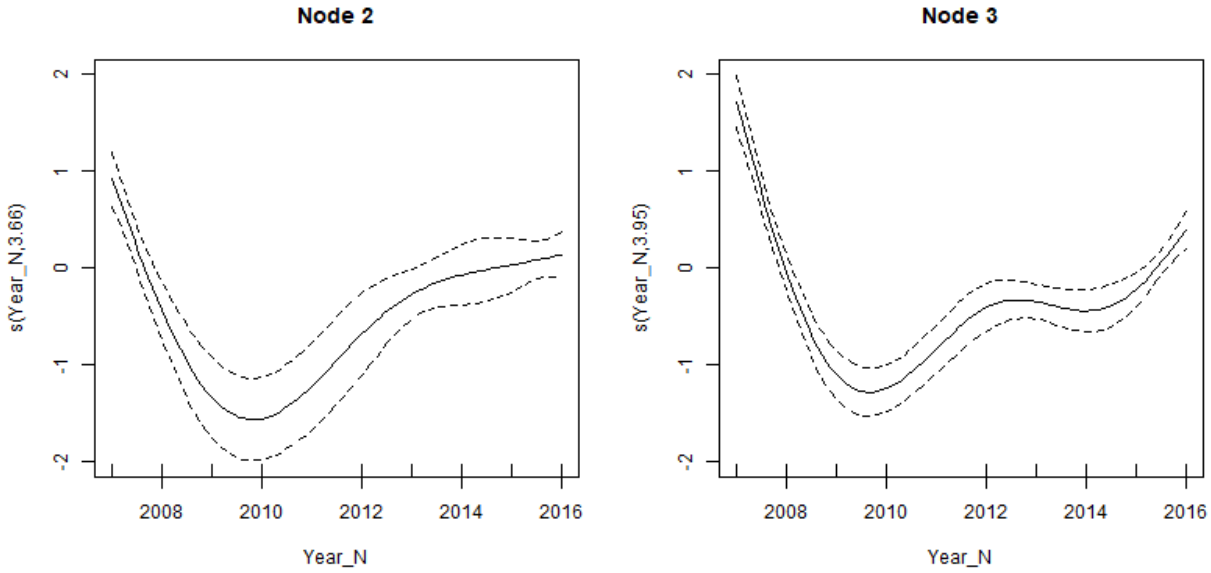

Figure 36: Smooth term fit estimated by gamma model to positive biomass (>0) data in each subtidal node.

2c. Tweedie model

The estimated smooth fits from the Tweedie model (Figure 37) are very similar to those from the binomial model.

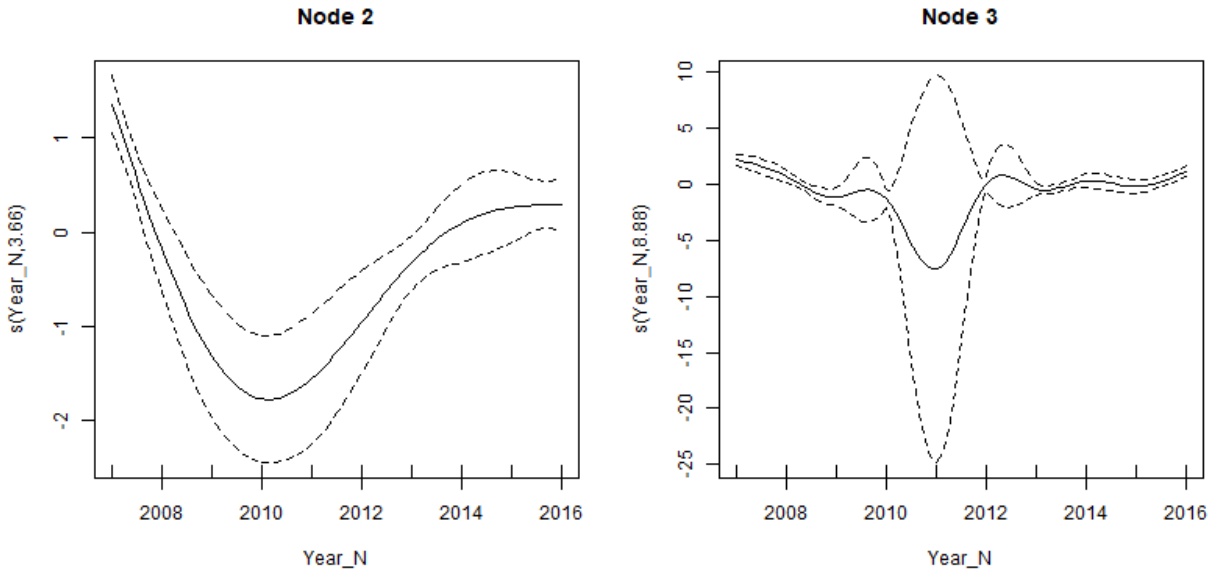

Figure 37: Smooth term fit estimated by Tweedie model to biomass data in each subtidal node for subtidal data. Note that the two Figures have a different y-scale.

2d. Tweedie estimated trend and uncertainty for each node

The estimated means and 95% confidence intervals for Node 3 are slightly higher and more variable than Node 2, but the overall patterns are similar (Figure 38). There was insufficient data to calculate confidence intervals in 2007 and 2011 in Node 3.

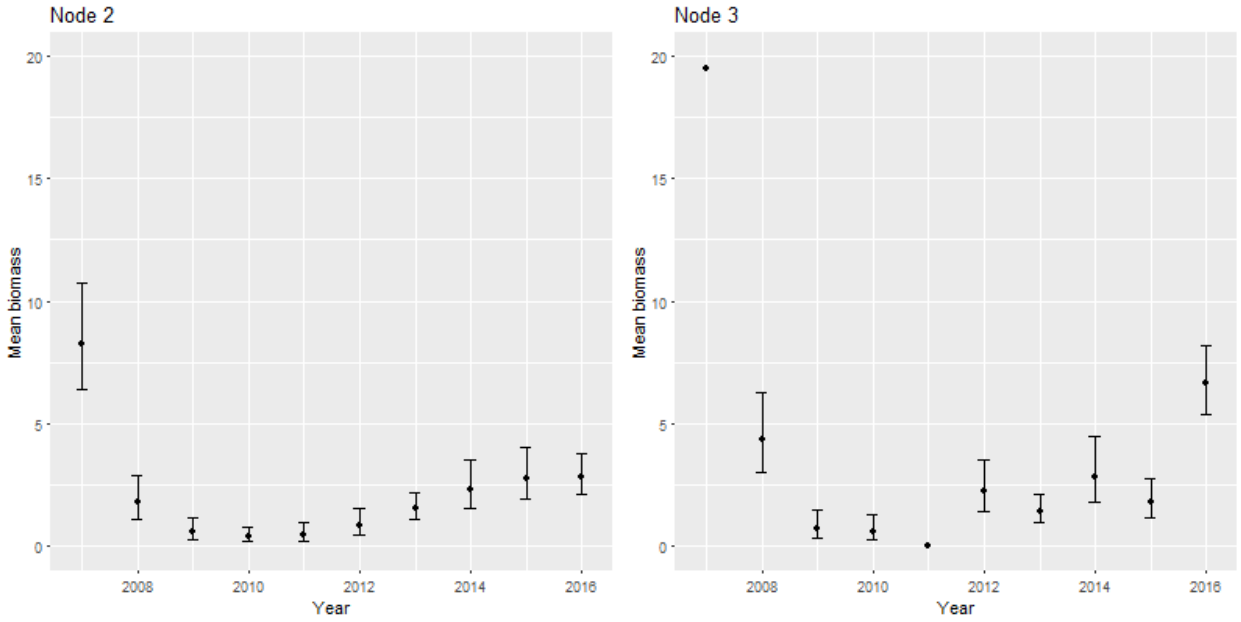

Figure 38: Estimated mean and 95% confidence intervals based on the Tweedie model for each subtidal node

Discussion

The regression trees based on the single species showed differing degrees of accuracy, with each species resulting in a different tree (although some splits were common). Including latitude and longitude improved the explanatory power of some of the tree's, however the inclusion of latitude and longitude restricts the application of the model to this location and provides less information about the environmental conditions driving the patterns in species abundances. To allow the relationships between species' biomass and environmental drivers to determine the tree splits, latitude and longitude were removed in all subsequent models. However, we note that the importance of latitude and longitude for some species may be an indication that some underlying environmental process are not otherwise captured in our existing covariates.

We explored a range of MRTs in Cleveland Bay to determine the most pertinent to the desired-state context. The first trees were based on a response matrix of the square root transformed biomass of all seagrass species. To ensure future relevancy for reporting we separated the data into Intertidal and Subtidal and fit separate trees to each so that the final nodes could be attributed to the habitat categories as determined in Udy et al. 2018. The trees were robust to the inclusion/exclusion of the data not collected as part of the long-term monitoring program. Analysing the data from each year separately resulted in community assemblage types that were similar from year to year in the years of higher mean biomass. In years of lower seagrass biomass there was less diversity in the community types but the patterns in assemblages were fairly consistent. Given we are interested in quantifying the desired state, which would by definition be a state based on years of high seagrass abundance, we excluded the years 2009 to 2012 from the remainder of the tree analyses.

The final trees we fitted were based on a presence-absence response matrix. Moving to a presence-absence response prevents species with higher biomass from dominating the analyses and instead weights all species equally. The results were very similar to the transformed biomass response, demonstrating that the community type classifications are fairly insensitive to transformations of the response. We recommend presence-absence as the most appropriate response for this analysis as the mere presence of some of the less dominant species may provide a good indication of a different community type. The recommended analyses resulted in nine community types in Cleveland Bay. If for some reason it was decided that some of the community types are so similar they should be pooled, the tree provides a mechanism for doing this, by stopping splitting higher up the tree.

Sediment type was important in differentiating community type in many of the tree analyses undertaken. This variable is not routinely collected meaning that it may not be possible to distinguish between community types in areas where the type of sediment is unknown. Management may wish to consider adding the collection of this data as part of the routine monitoring program, facilitating the possibility of predicting community types in areas where little biological data is available.

Once the species communities were established using the MRT method, we modelled the total biomass in each node to determine the trends in the different community types. This gives us a mean biomass and associated confidence interval for each community type for a time-series of 11 years. Over the study period the mean biomass of seagrass fluctuated and showed significant loss and subsequent recovery. Through further examination of these temporal trends the broader NESP project team will be able to set desired state targets for seagrass for the different community types in Cleveland Bay. We expect the desired state

targets will be higher for some community types due to their higher mean biomass when operating in a “usual” or “recovered” state.

While this analysis has focused entirely on Cleveland Bay where there is good information on the biomass of seagrass both spatially and temporally, we have ensured that the methodology used could equally be applied to other areas with sufficient data. If the analysis were to be completed on a broader area with more differentiation between sites, data permitting, then it would be beneficial to add further explanatory variables such as climatic/pressures measures to help explain the differences in temporal trends. This would assist in establishing what the desired state may be under ‘regular’ climatic conditions.

There are other assemble and model methods that we could have used to complete this analysis. The methods used were discussed and chosen during a Reef 2050 seagrass expert working group meeting. The results are shown to be robust to the assumptions and biologically defensible. However, we expect that choosing a different combination of methods may result in small changes in the community composition analysis.

References

Carter, A.B., McKenna, S.A., Rasheed, M.A., McKenzie, L., Coles, R.G. 2016, Seagrass mapping synthesis: A resource for coastal management in the Great Barrier Reef World Heritage Area. Report to the National Environmental Science Programme. Reef and Rainforest Research Centre Limited, Cairns (22 pp).

Clark, L.A., Pregibon, D. 1992, Tree-based models. Pages 377-420 in J.M, Chambers and T.J. Hastie, editors. Statistical models in S. Wadsworth and Brooks, Pacific Grove, California, USA.

Davies, J.N., Rasheed, M.A. 2016, Port of Townsville Annual Seagrass Monitoring: September 2015, James Cook University Publication, Centre for Tropical Water & Aquatic Ecosystem Research (TropWATER), Cairns, 43 pp.

De'ath, G. 2002, Multivariate Regression Trees: A new technique for modelling species-environment relationships. Ecology 83 (4), 1105-1117.

De'ath, G. 2014, mvpart: Multivariate partitioning. R package version 1.6-2. https://CRAN.R-project.org/package=mvpart

Devlin, M. da Silva, E.T., Petus, C., Tracey, D. 2017, Marine Monitoring Program: Annual report for flood plumes and extreme weather 2013-2014. Tropical Water & Aquatic Ecosystem Research (TropWATER) Publication. Report No 15/63. Great Barrier Reef Marine Park Authority, Townsville, Australia. 102 pp.

Hastie, T.J., Tribshirani, R.J., Friedman, J. 2009, The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Second Edition. New York: Springer-Verlag.

Larsen, D.R., Speckman, P.R. 2004, Multivariate Regression Trees for analysis of abundance data*, Biometrics* 60, 543-549.

Mullahy, J. 1986, Specification and Testing of Some Modified Count Data Models*, Journal of Econometrics*, 33, 341-365.

R Core Team (2017). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Therneau, T., Atkinson, B. and Ripley, B. 2017, rpart: Recursive Partitioning and RegressionTrees. R package version 4.1-11. https://CRAN.R-project.org/package=rpart

Tweedie, M. C. K. 1984, An Index Which Distinguishes between Some Important Exponential Families, in J. K. Ghosh and J. Roy, eds., *Statistics: Applications and New Directions—Proceedings of the Indian Statistical Institute Golden Jubilee International Conference*, 579–604.

Udy, J., et al. 2018 (In prep), Monitoring seagrass within the Reef 2050 Integrated Monitoring and Reporting Program: Draft Final report of the seagrass expert group. Report for the Great Barrier Reef Marine Park Authority, Great Barrier Reef Marine Park Authority, Townsville.

Wells, J.N, and Rasheed, M.A. 2017, 'Port of Townsville Annual Seagrass Monitoring and Baseline Survey: September - October 2016', James Cook University Publication, Centre for Tropical Water & Aquatic Ecosystem Research (TropWATER), Cairns, 54 pp.

Wood, S.N. 2017. Generalized Additive Models: An Introduction with R (2nd edition). Chapman and Hall/CRC

Appendix A. Regression trees for individual species

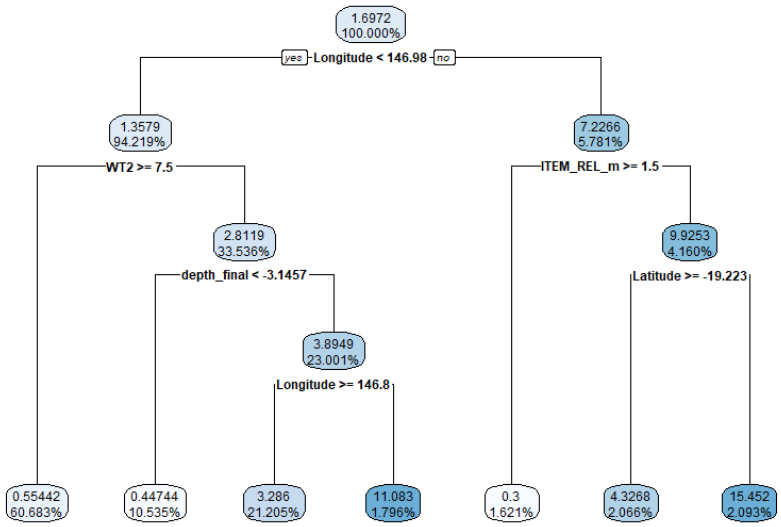

Figure A1: Regression tree for *Cymodocea serrulata* (CS), including latitude and longitude as predictor variables. Relative error =0.9092

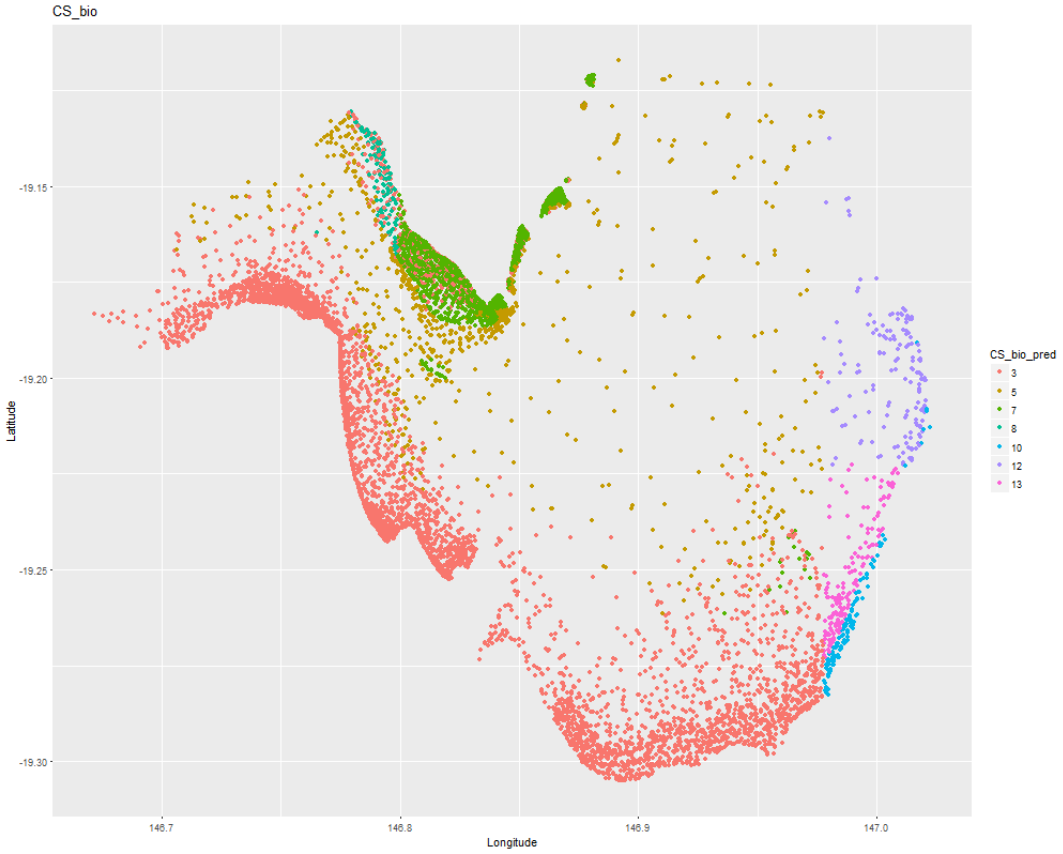

Figure A2: Map of predicted nodes for *Cymodocea serrulata*, including latitude and longitude as predictor variables. Note: removing latitude and longitude resulted in no splits i.e there is no tree for *Cymodocea serrulata* without latitude and longitude.

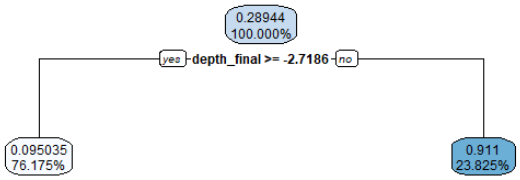

Figure A3: Regression tree for *Halophila spinulosa* (HS), including latitude and longitude as predictor variables. Note: latitude and longitude are not selected so the tree is the same regardless of whether they are included or not. Relative error =0.9807.

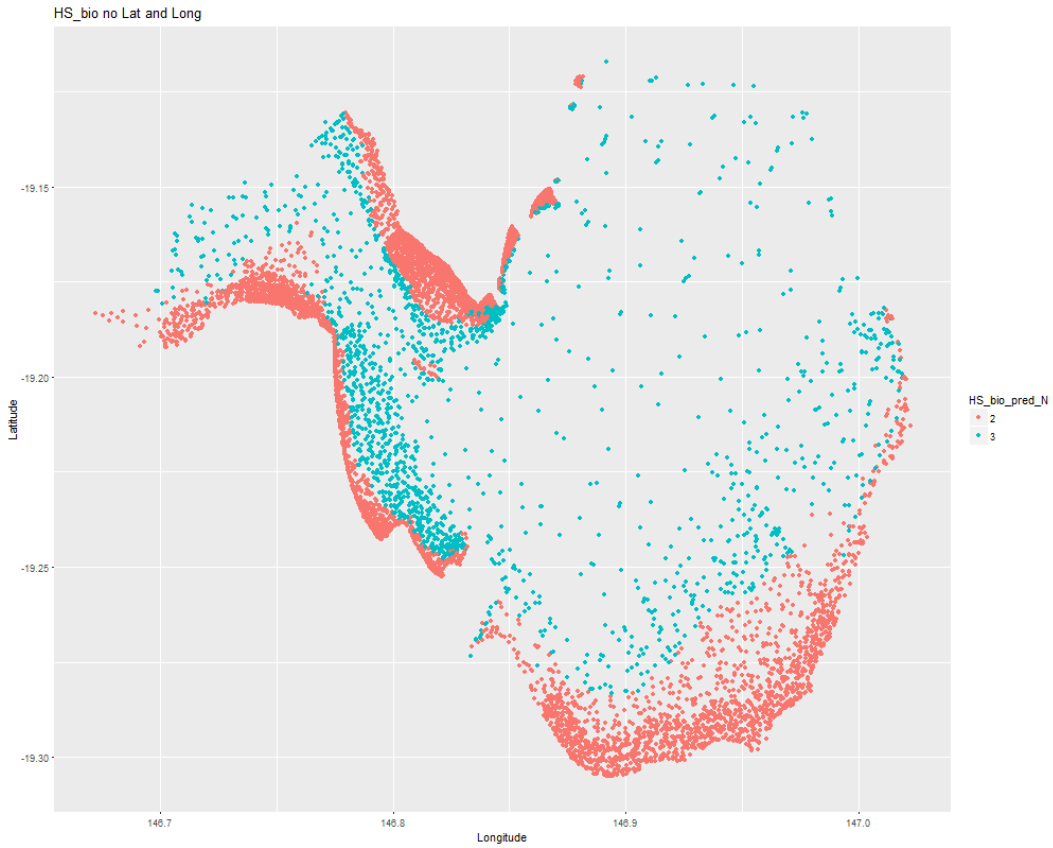

Figure A4:` Map of predicted nodes for *Halophila spinulosa*, including latitude and longitude as predictor variables

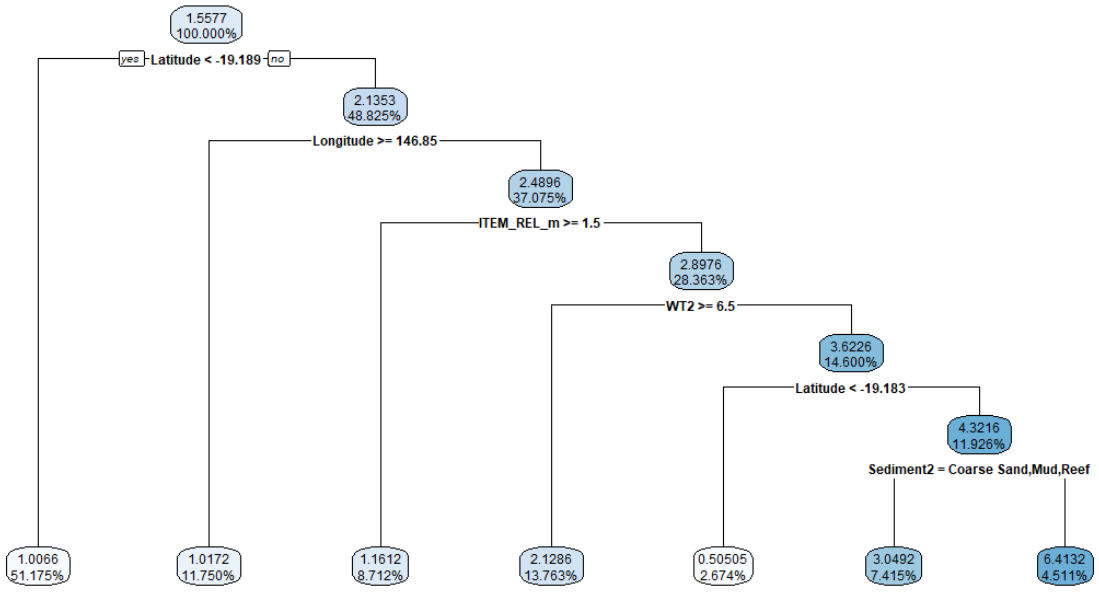

Figure A5: Regression tree for *Halodule uninervis* (HU), including latitude and longitude as predictor variables. Relative error =0.9269.

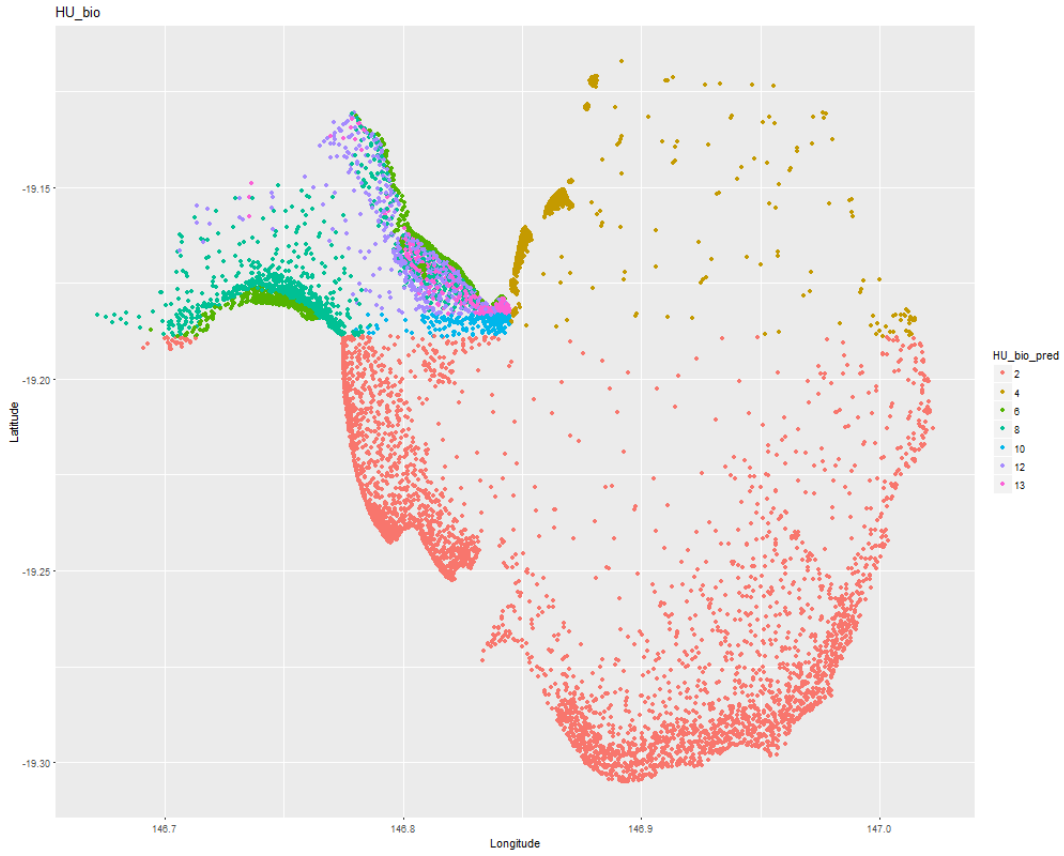

Figure A6: Map of predicted nodes for *Halodule uninervis*, including latitude and longitude as predictor variables

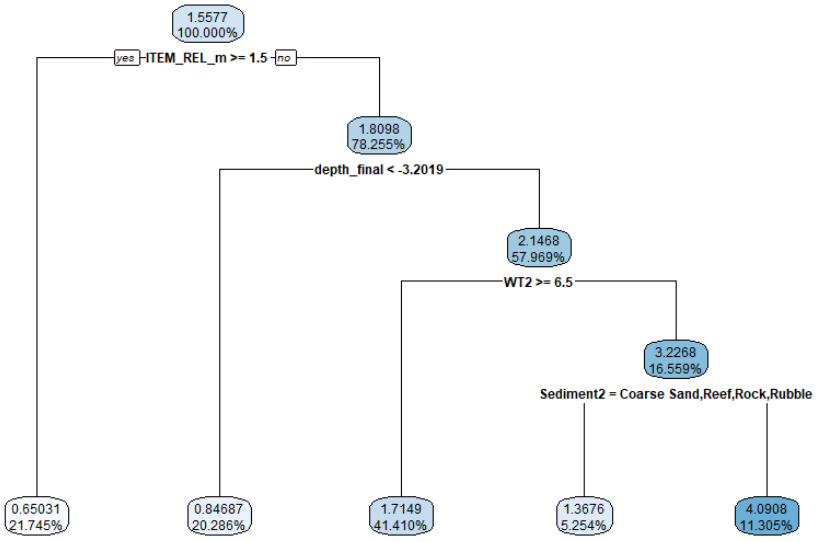

Figure A7: Regression tree for *Halodule uninervis*, without latitude and longitude as predictor variables. Relative error =0.9505.

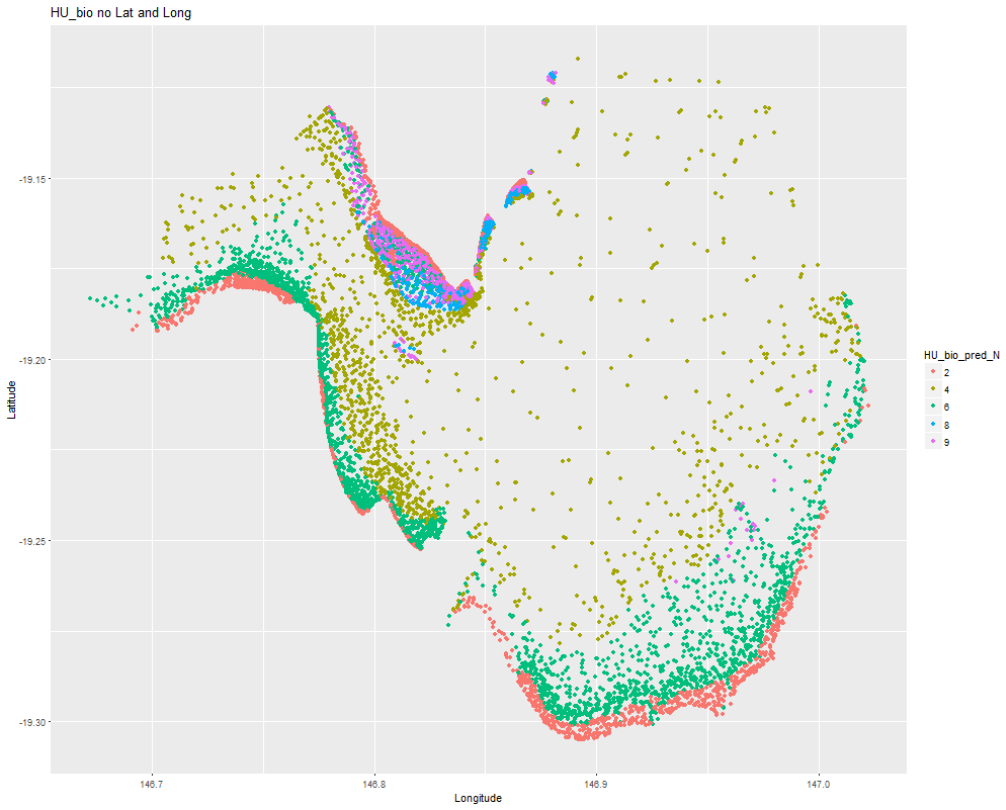

Figure A8: Map of predicted nodes for *Halodule uninervis*, without latitude and longitude as predictor variables

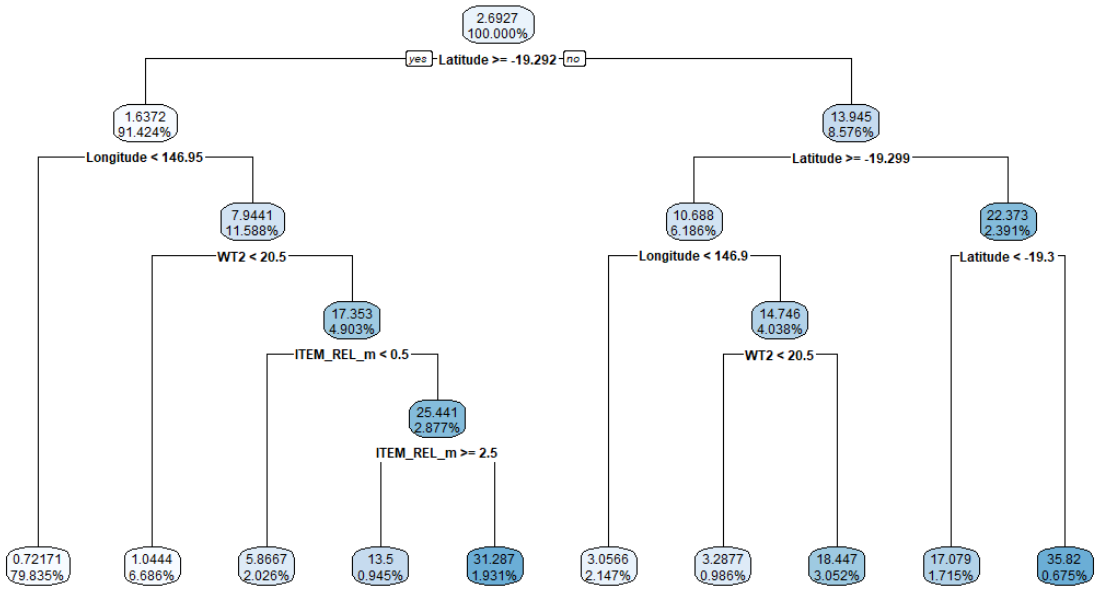

Figure A9: Regression tree for *Zostera muelleri subsp capricorni* (ZC), including latitude and longitude as predictor variables. Relative error =0.7159.

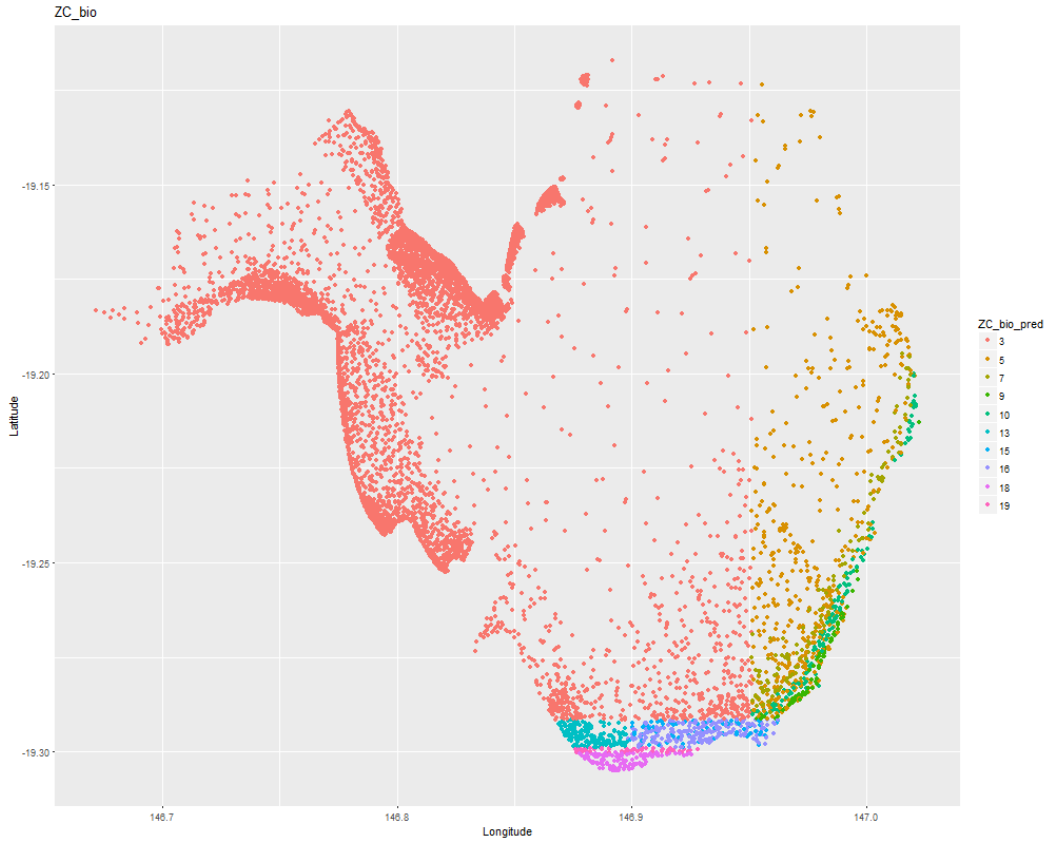

Figure A10: Map of predicted nodes for *Zostera*, including latitude and longitude as predictor variables

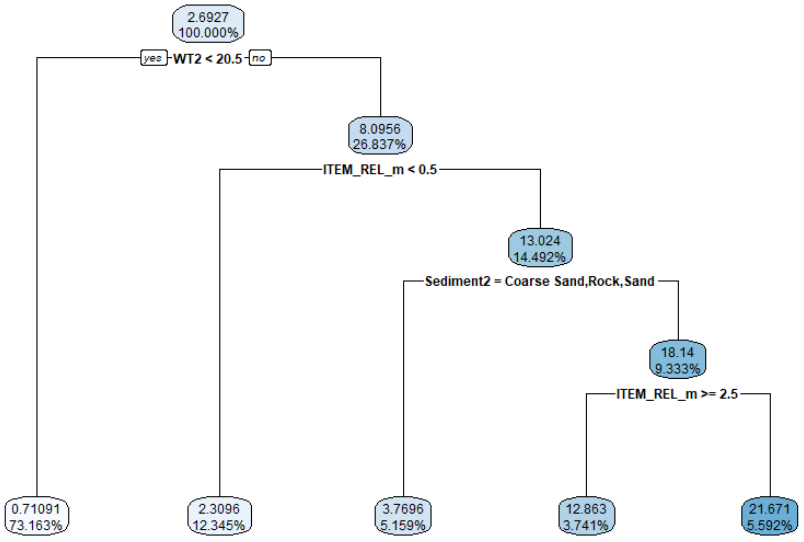

Figure A11: Regression tree for *Zostera* (ZC), without latitude and longitude as predictor variables. Relative error=0.8027.

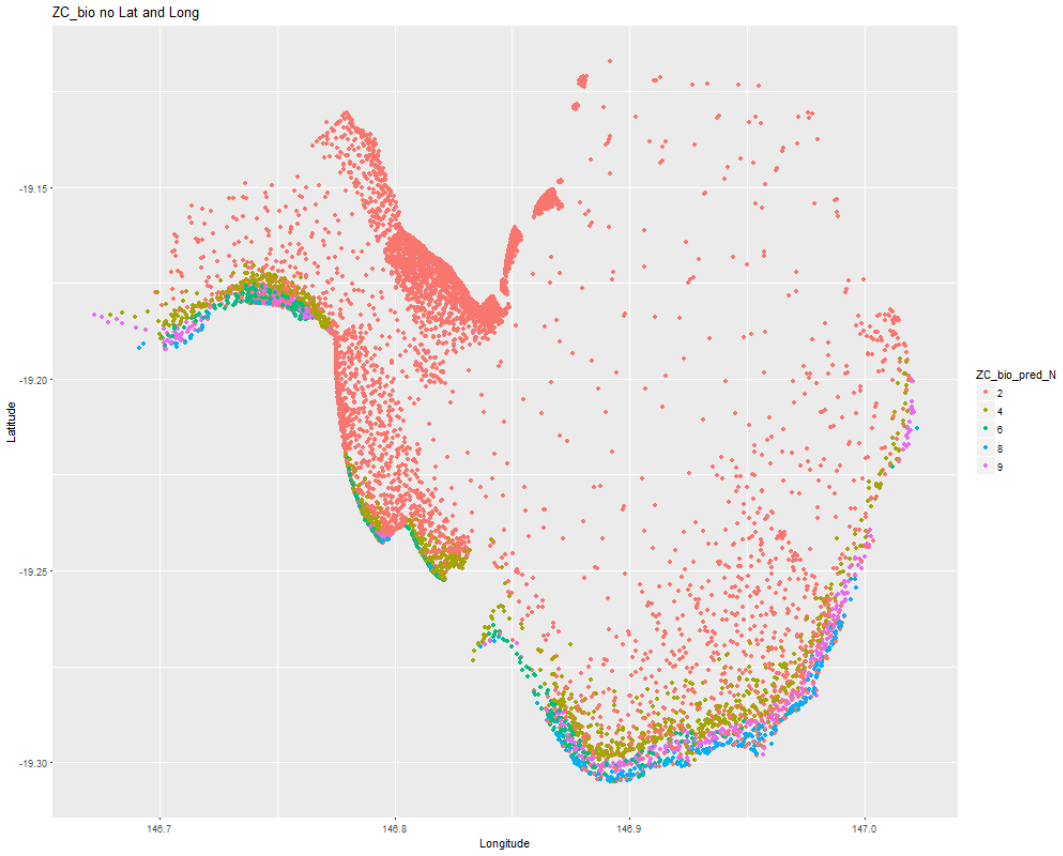

Figure A12: Map of predicted nodes for *Zostera*, without latitude and longitude as predictor variables

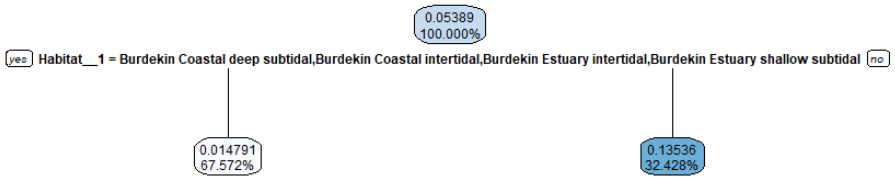

Figure A13: Regression tree for *Halophila decipiens* (HD). Note latitude and longitude were not selected as predictor variables. Relative error =0.9842.

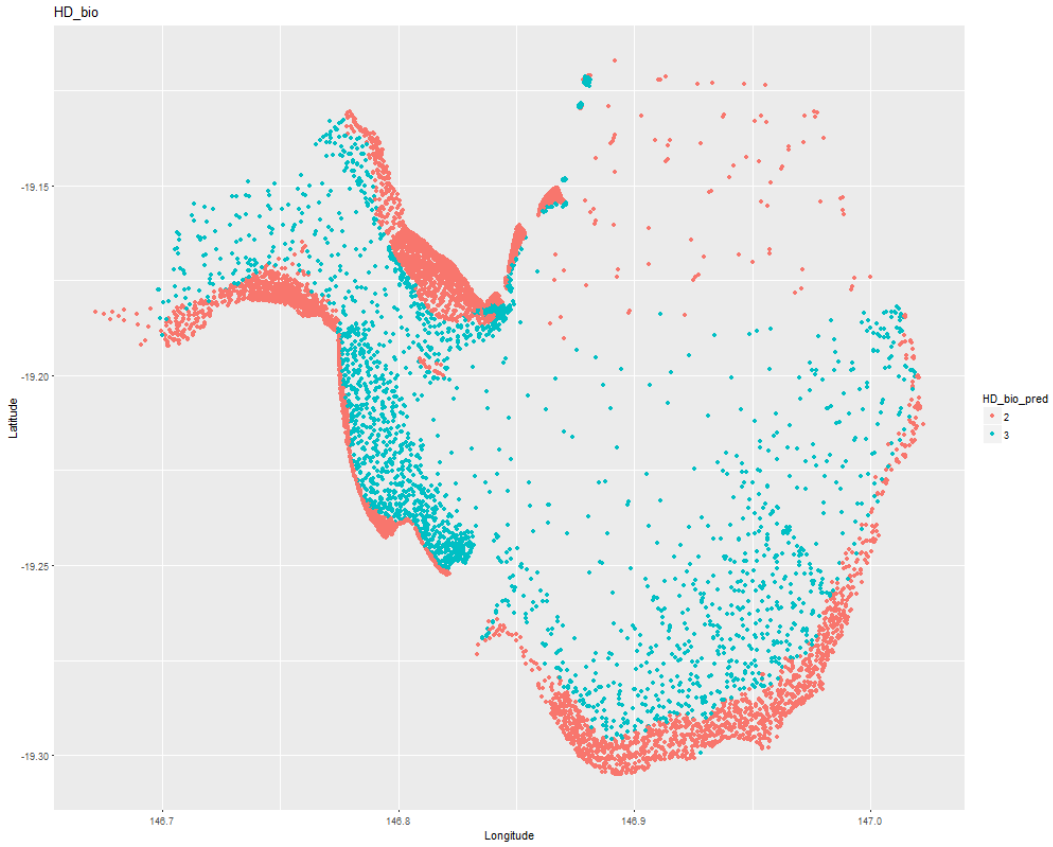

Figure A14: Map of predicted nodes for *Halophila decipiens*, including latitude and longitude as predictor variables

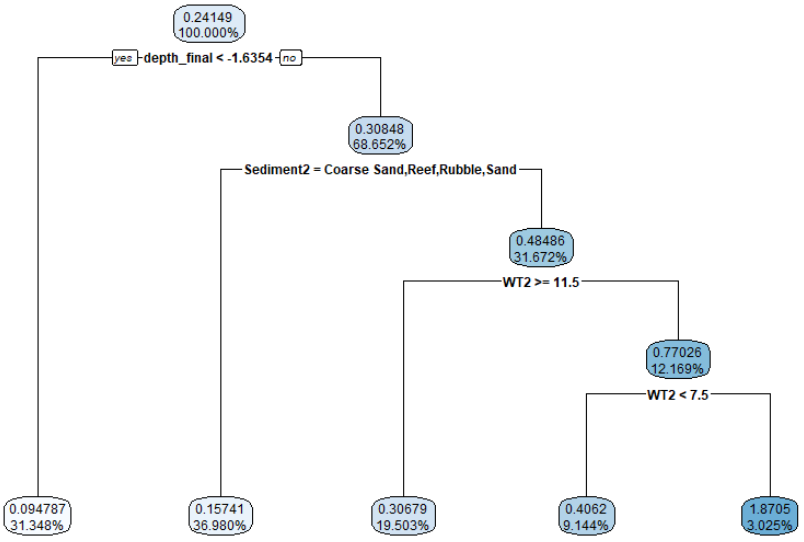

Figure A15: Regression tree for *Halophila ovaris* (HO) including latitude and longitude as predictor variables. Relative error =0.8146.

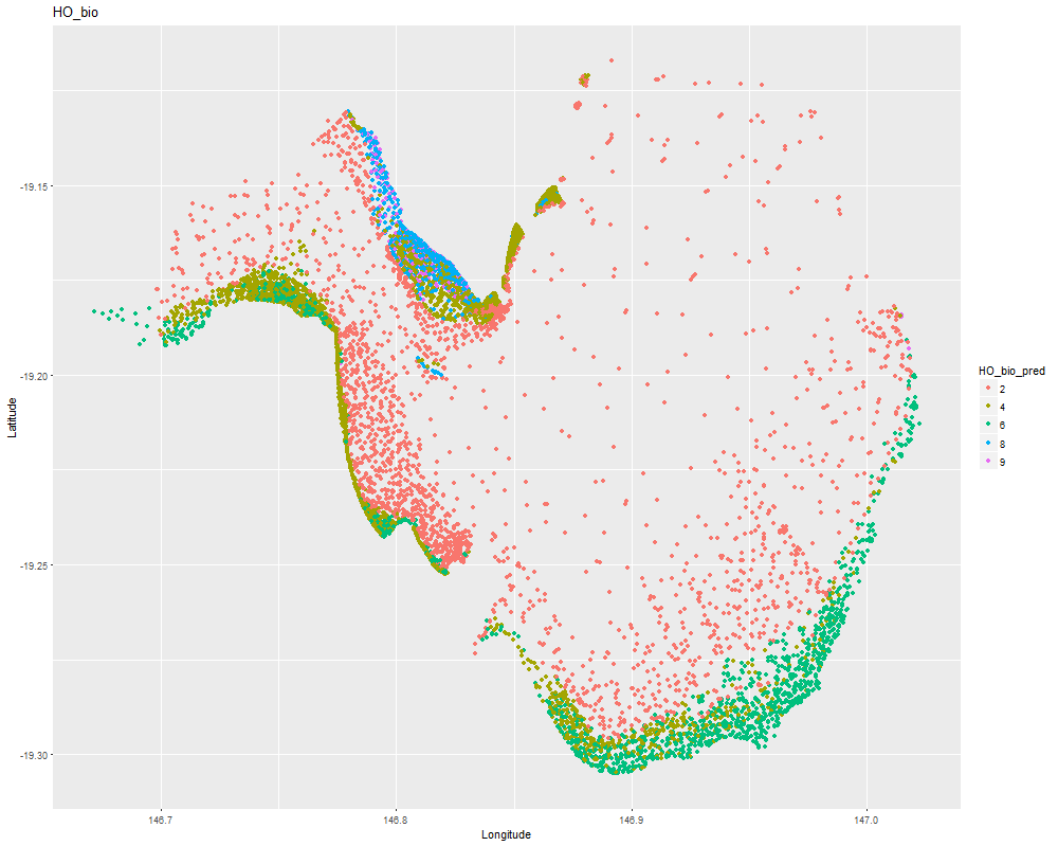

Figure A16: Map of predicted nodes for *Halophila ovaris*, including latitude and longitude as predictor variables