MSC DATA SCIENCE & ANALYTICS DISSERTATION

Standardized Water Level Index Reconstruction

Author :

Maithri Amarnath Vijayakumari

(19251022)

Supervisor:

Dr Niamh Cahill

A thesis submitted in fulfilment of the requirements

for the degree of Master's in Data Science and Analytics

2019-2020

to the

Department of Mathematics & Statistics

Maynooth University

Acknowledgement

I would like to first acknowledge and thank my dissertation supervisor Dr Niamh Cahill of the Mathematics and Statistics department at Maynooth University. Her words of constant encouragement & meticulous reviews helped me to improve the quality of my thesis work in many aspects. She consistently steered me in the right direction whenever she thought I needed it.

I would also like to thank Dr Catherine Hurley Head of Department of Mathematics and Statistics at Maynooth University for the support and learning offered throughout the year.

Maynooth University
National University of Ireland Maynooth

Department of Mathematics & Statistics

Abstract

"Standardised Water level Index Reconstruction"

by Maithri Amarnath Vijayakumari

The thesis work presents using assemblies of salt-marsh foraminifera as indicators at sea level, collected from eight different locations along the East and West Coastal region of United States, where their abundance is used as the key predictive variables of Standardized Water Level Index in this work. We implement Random Forest based on this species count of foraminifera to establish a predictive model for standardized water level index (SWLI) for each region of United State and it was found that Random Forest provided a better performed model with Mean Absolute Error ranging from 12.01 to 22.62 for 8 different regions across United States with 10-fold cross-validation on12 the dataset. Species clustering or grouping using PCA and clustering techniques such as K-means clustering, and hierarchical clustering was implemented to explore the impact of minimizing the number of species variables. Relationship between the grouped species and tidal elevation was quantified, by observing the characteristics of the clusters and by following a clustering technique implemented and measured by species abundance with respect to elevation. A standardized prediction of elevation was made by this clustered species for each region of United States. A comparison was carried out on the prediction model results developed with Random Forest on clustered and non-clustered data, where model with lower dimensional variables also resulted in better predictive SWLI with Mean Absolute Error ranging from 13.51 to 31.00 for 8 different regions across United States. Thus, resulting in usage of lower-dimensional variables for the larger dataset in real practice for reconstruction of the Standardized Water level Index.

Keywords: Random Forest, Clustering, PCA, K-means Clustering and Hierarchical Clustering.

Table of Contents

List of Tables

List of Figures

1. Introduction

The variety of technologies widely developed helps us to determine the sea level with tide gauges every couple of minutes and by satellite across whole oceans. Nevertheless, this may only give a small portion of the broader picture. We ought to aggregate the proof over a broader range of periods for a detailed evaluation, from minutes to decades or perhaps even longer [1]. The "International Geological Correlation Program (IGCP) project- 61 (Sea-level movements during the last deglacial hemicycle) and project- 200 (Late Quaternary sea-level changes: measurement, correlation, and future applications)" proposed a widely applicable approach to reconstructing relative history at sea level for various locations and ecosystems [2]. Significant predictors of historical sea-level changes comparative to the present, with quantitative uncertainty terms, come from situ sediments, fossil animals, salt marshes, morphological and archaeological features influenced by the paleo sea level. Each of these features represents an example of an indicator at sea level [3].

Salt marshes are marine habitats within a fairly small range of salinity and tidal range. Across the ecology of both marine and terrestrial environments, salt marshes are vital connections. Foraminifera species found in high abundance in salt marsh sediments alter their existence in terms of tide elevation [4]. The relevance of species distribution is so prominent that elevation is also considered the controlling factor in tidal ranges [4]. This paper articulates that using fossilized assemblies of salt-marsh foraminifera as indicators at sea level are key predictive variables of Standardized Water Level Index.

Foraminifera are unicellular microstructural species that are hard coats and live in a variety of aquatic environments. Due to their richness and relatively high abundance, combined with their vulnerability to different environmental variables (for example, temperature and salinity), geoscientists have been able to recreate the ocean's past conditions [4]. While the fact that foraminiferal experiments in marsh wetlands abundantly rendered them ideal for statistical analysis was noted by Scott and Medioli (1978), much of the early research on tidal elevation assemblage was focused on visual data assessment [4]. More recent work appeared to use statistical techniques to describe co-relationships and predictions between foraminifera assemblages and elevation of tide.

One obvious interpretation is that a foraminiferous distribution might not be the only "environmental aspect" in determining tidal elevation. This is not surprising as the distribution of species is the result of many complicated interactions between organisms and their environment. Although spatial variation has been observed in previous studies by Horton and Edwards [5] suggesting that tidal elevation is also one of the prominent control variables, the strong correlation between

foraminiferous species and elevations reflects the wide environmental gradient across the coastal region [4]. High elevation is thus a substitute of one of the influencing variables. In the later sections of this paper its strong correlation with the distribution of species is quite reasonably explained.

In the late 1970s and early 1980s, David Scott and colleagues published seminal articles on the use of foraminifera as precise standardized water level indicators. This initial work focused on the finding that distinct assemblages of saltmarsh foraminifera appeared to characterize various elevation ranges within the coastal area [4]. Although initially researching based on a small number of saltmarshes, ample empirical evidence now exists that distinct marine foraminiferal assemblages can be associated with tidal elevation. The study reported similar foraminifera assemblage's data in salt marshes for the analysis that assemblages are predictive variables to tidal elevation.

In this thesis we implement Random Forest [6], based on the presence and abundance of certain species of foraminifera in salt marshes of the specified dataset, to establish a predictive model for standardized water level index (SWLI) at different locations in the United States. SWLI is expressed as standardized water level elevation or tidal elevation due to the difference in this tidal range among different sites of United regions. Foraminiferous accumulation can be used as a valid indicator of SWLI as mentioned above, as the salt marshes represent the tide period.

To explore the impact of minimizing the number of species variables, "assemblage grouping" or "species clustering" is explored using PCA [7] and clustering techniques such as K-means clustering [8] [9] and hierarchical clustering [9]. The emphasis for clustering is placed on the distributions of the individual species that comprise the assemblage of interest with respect to elevation. If the relationship between the grouped species and tidal elevation can be quantified, by observing the characteristics of the groups, measured in some way by their abundances with respect to elevation, a standardized calculation of elevation can be made by that grouped assemblages.

The rest of this study is organized as follows: first, an overview of the data, variables used in the data set respectively, few exploratory visualizations. Then a detailed description of the methods such as Implementing PCA [7], K- means Clustering [8] and Hierarchical Clustering [9] and their analysis to achieve the results. The conclusions are illustrated in later chapters preceded by the discussion.

2. Data Exploration

An overview of the variables used in the datasets developed for standardized water level index prediction is illustrated in Table 2.1.1. Salt marshes are of low biological diversity, with few species adapted to conditions created by high environmental gradients and many species living in salt marshes are specialists.

Table 2.1.1: Description of the variables in dataset created for prediction of SWLI.

VARIABLES	DESCRIPTION
Coast	*Coastal Regions of the United States namely – East & West Coasts.*
Region	*Regions across the United States namely – California, Oregon, Connecticut, Maine, New Jersey, Newfound Land, Vancouver Island & Washington*
Site	*Samples collected at 36 sites of different regions of East Coast and 19 Sites of different regions of West Coast, respectively.*
Sample ID	*Unique IDs are given for each sample collected at different sites of East and West Coastal regions of the United States.*
SWLI	Standardized water level index or the tidal elevation measure of seawater.
Species	*Species identified from samples collected are classified based on their respective scientific names namely- Jm.Bp, Ti.Sl, Tc, Hs, Mf, Pl, Pi, Am, Ai, Ea, Tq, Mp, Calc, Rs, Ab, As, Ts, Ad, TO¸ PH, TG¸ Tt, Ph.*
Abundance	*Counts of each species recorded from each sample collected.*

These salt marsh distributions have been collected from the coastal regions of the United States, namely the West Coast and the East Coast. Specific regions were chosen to collect samples of these salt marshes along the coastlines of the United States, namely California, Oregon, Connecticut, Maine, New Jersey, Newfound Land, Vancouver Island and Washington. Samples [10] [11] [12] collected were

provided with a unique recognition ID as illustrated in Table 2.1.1. Along 36 sites of the East Coastal regions of the United States and 19 sites of different regions of the West Coast where these distinct salt marsh foraminifera assemblies were gathered. Species from 5 different published [12] [11] [10] cores, namely West Coast Core, NFLD Core, CT Core, NJ Core and Maine Core, have also been provided for this dissertation. The corresponding SWLI must be predicted for those 5 locations where the counts of this foraminifera have been recorded.

Foraminifera now occupies a central place in this study evaluating the significance of variations in Standardised water level index, since they can bridge the difference between short records of instruments and long-term geological data [4]. To order to obtain accurate data conducive to statistical analysis, a certain minimum number of Foraminiferal must be counted. The precise figure needs to be the relative abundance to saltmarshes of different taxa in regions of the United States.

Table 2.1.2: Taxonomy of the species of foraminifera identified in the samples collected across US.

Species	Taxonomy of Foraminifera
Jm.Bp	*Jadammina macrescens + Balticammina pseudomacrescens*
Ti.Sl	*Trochammina inflata + Siphotrochammina lobata*
Tc	*Tiphotrocha comprimata*
Hs	*Haplophragmoides spp.*
Mf	*Miliammina fusca*
Pl	*Pseudothurammina limnetis*
Pi	*Polysacharimma ipohalina*
Am	*Arenoparrella mexicana*
Ai	*Ammoastuta inepta*
Ea	*Eggerella advena*
Tq	*Trochammina squamata*
Mp	*Miliammina petila*
Calc	*All calacreous species*
Rs	*Reophax spp.*
Ab	*Ammobaculties spp.*
As	*Ammotium spp.*
Ts	*Textularia sp.*
Ad	*Ammodiscus spp.*
TO	*Trochammina ochracea*
PH	*Polysaccammina hyperhalina*
Tt	*Trochamminita spp*
TG	*-Unknow*
Ph	*Paratrochammina haynesi*

The use of different taxonomies abbreviations in sea-level foraminifera-based on this study creates additional complications and ambiguity when reading literature. Table 2.1.2 is the brief guide for the foraminifera species taxonomy based on some of the key species included in this dataset.

Specific assemblies of the foraminifera exist on different or same Standardized Water levels. The study of this foraminifer in surface sediment allows quantification of relationships between species and elevation, which can be reasonably accurate. Based on this information, the tidal elevation or SWLI can be estimated, at which a species accumulated from its foraminifera.

2.1 Variation of Species w.r.t Elevation

Species diversity and abundance vary across the elevation as shown in the ggvis [13] plot in Figure 2.1.1. And considering the few species, we can interpret that species such as Jm.Bp, Ti.Sl, Tq, Mp, Tt are present in abundance from elevation 1.4 to 2.6. While species Mf, Ab, etc are present in the range of elevation from 0.6 to 2.0.

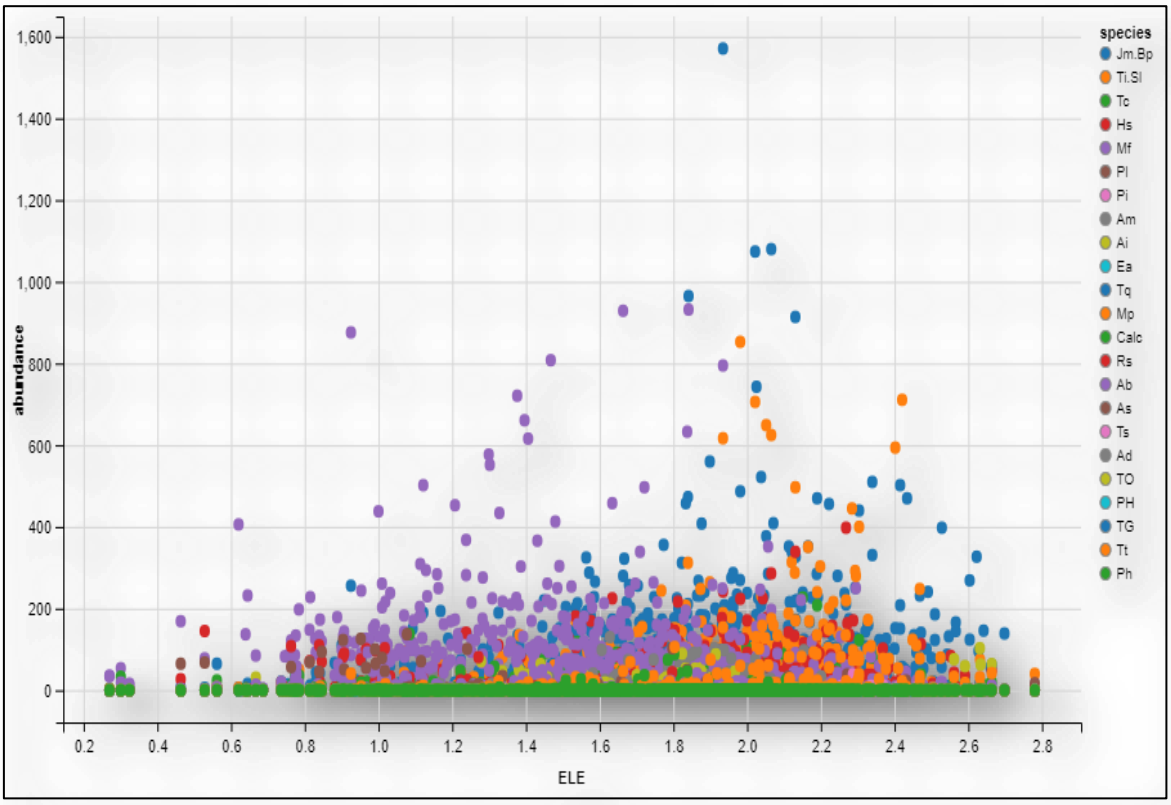

Figure 2.1.1: Species variation across all regions of United States with respect to Elevation.

2.2 Variation of Species w.r.t Elevation across regions of US

Species of foraminifera variation and elevation correlation abundance can be interpreted using the ggvis plot [13] in detail by comparing regions where salt marsh samples have been accumulated from several coastal sites in the United States.

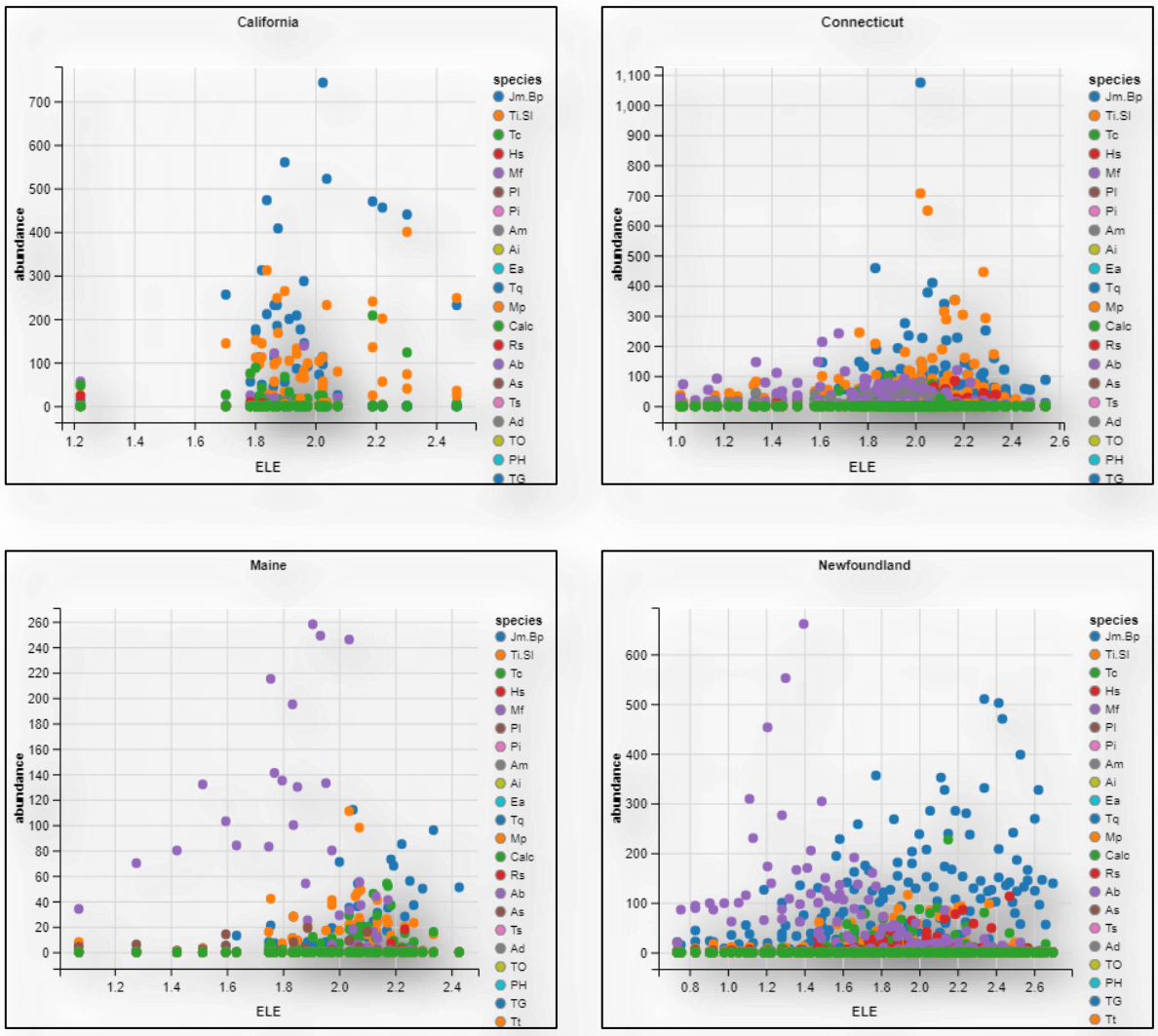

Figure 2.2.1: Species variation across California, Connecticut, Maine & Newfound Land regions of US with respect to Elevation.

California and Maine have a smaller diversity of species compared to the other two regions of the United States, namely Newfound Land and Connecticut, as shown in Figure 2.2.1. Species such as Jm.Bp, Tq & TG are abundant and are gradually increasing from the lower (0.8) to higher (2.6) elevations of Newfound Land. While species like Ti.Sl & Mp are abundant from mid-range(1.8) elevation to higher (2.4) elevations in California Maine & Connecticut.

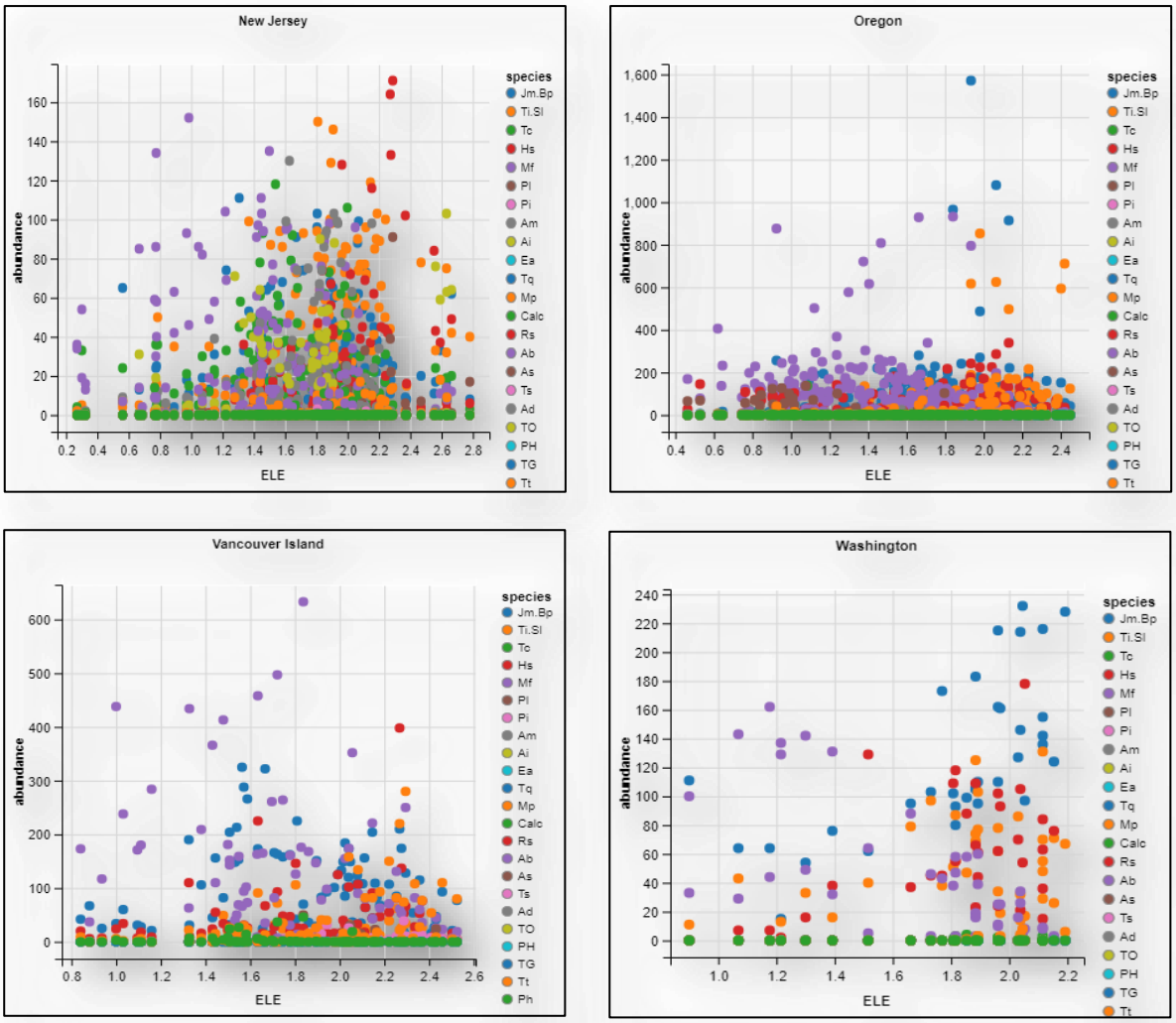

Figure 2.2.2: Species variation across New Jersey, Oregon, Vancouver Island & Washington regions of US with respect to Elevation.

The same features as Figure 2.2.1 are shown in Figure 2.2.2 which shows foraminiferal distributions of salt marshes from the east coast of the United States to New Jersey, Vancouver Island and Washington. In Figure 2.2.2 New Jersey clearly demonstrates that species are diversely available and large in number in the respective region compared to others. While species such as Mf & Ab exist from lower(0.4) to middle(1.8) range elevations in regions of Oregon and Vancouver Island. Washington has only traces of certain species, namely Jm.Bp, Tq, TG, Ti.Sl, Mp, Hs and Rs.

2.3 Samples of foraminifera at different levels of SWLI

The density plot was constructed using the ggridges package [14] and viridis package [15] to analyse that more samples were recorded in the East Coast region of the United States at 180 to 230 SWLI values. While samples from West Coast regions are recorded from 140 to 230 SWLI values as shown in Figure 2.3.1.

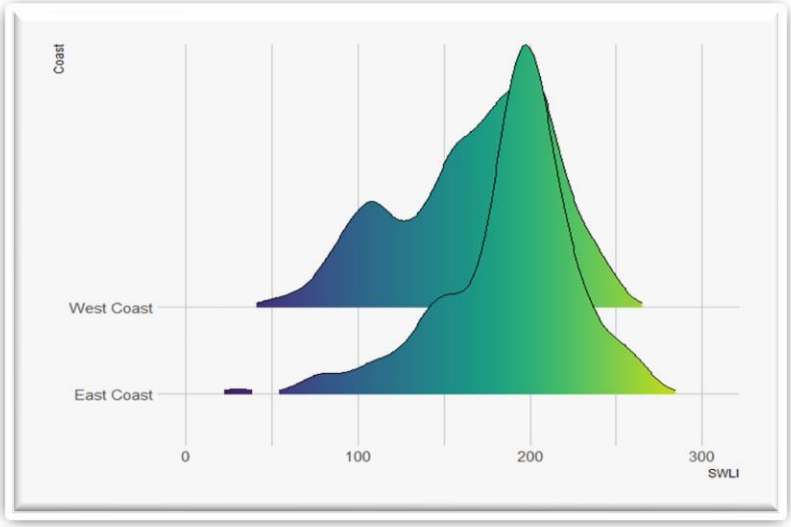

Figure 2.3.1: Number of Samples collected across East and West Coastal Regions of US

A similar plot was constructed to visualize across regions and Figure 2.3.2 shows that California, Connecticut, Washington, and Main had a higher number of samples recorded at 170 to 230 SWLI values. While New Jersey, Oregon, Newfound Land and Vancouver Island showed a gradual trend from 100 to 250 SWLI values.

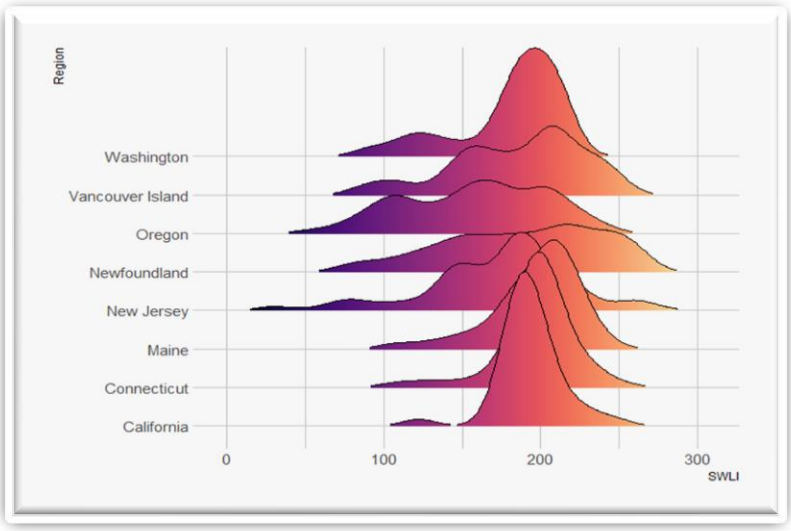

Figure 2.3.2: Number of Samples collected across Regions of US at different levels of SWLI

3 Methods

This portion of this study focuses on the methods studied and carried out in the reconstruction of the United States coastal region's standardized water level index. The analysis of the effect of minimizing twenty-three species count variables is evaluated by using PCA and "species grouping" by clustering techniques such as K-means clustering and Hierarchical clustering A predictive model for SWLI is established and analysed by applying Random forest on the data without clustering the species variables and on clustered species dataset.

3.1 Method - I: Implementing PCA

Principal Component Analysis [7] is one of the most important methods to minimize dimensionality in the extraction of efficient components from large dimensional input data vectors [16]. PCA is a process in which significant variables (in the component form) derive from a large number of variables that are available in a data set. A limited number of features of a high-dimensional data set are used in order to obtain as much information as possible to represent insignificant dimensions. And our datasets have more than three variables; visualizing and interpreting outcomes can be very difficult.

PCA [7] is used to derive useful data from this multivariate dataset as a set of specific new variables, called the principal components. These modern variables reflect a linear combination of the originals. PCA reduces the dimensionality of our multivariate dataset to two or three PCs, graphically visualized with minimal data loss. [17]. In this section, we will see the impact of PCA on decreasing the number of predictor variables for SWLI Reconstruction data consisting of twenty-three species counts to make reconstruction more efficient.

- The PCA was performed for each region of the United States.
- The basic R function prcomp() [18] is used to perform the PCA function. The variable is by default centred to have a mean equal to zero.
- Using prcomp() [18] function, we computed the standard deviation for each principal component. And in order to measure the variance ratio explained by each component, the variance is simply divided by the sum of the total variance.

We determine the number of components to be chosen for the modelling stage for each region of United States by plotting a scree plot which is shown in the later result sections.

3.2 Method - II: Implementing Clustering

Clustering [9] is another way of developing a technique for dimension reduction or grouping. "Clustering is the process of splitting the population or data points into a number of groups, in such a way that data points in the same groups are more comparable to other data points in the same category than others" [9]. It is essentially a set of objects which are focused on similarity and dissimilarities between them.

We interpreted the relationship of species with respect to elevation from the data exploration sections 2.1 (Variation of Species w.r.t Elevation) and 2.2 (Variation of Species w.r.t Elevation across regions of US). A clustering technique is built here using characteristics of the relationship between these species and elevation. And thus, clustering the species data to see if any clusters can help lower the number of twenty-three species variables in SWLI prediction.

3.2.1 Clustering technique

A clustering function is applied to raw species abundance data with the aim of grouping the species in such a way that the clustered species will maintain the ability to forecast elevation. Below are the steps which define the approach to clustering.

STEP – 1: This step presents the numerical summaries of the data to explain the distributions of the species with respect to elevation.

- The maximum abundance (max_ab), mean abundance (mean_ab), standard deviation for abundance (sd_ab) and abundance percentiles for each species in the dataset are calculated. Where columns "lower_95" and "upper_95" are the 2.5^{th} and 97.5^{th} percentiles for abundance, respectively. Similarly, "lower_68" and "upper_68" are the 16^{th} and 84^{th} percentiles for abundance, respectively. Table 3.2.1 illustrates an example of two species from the dataset.

Table 3.2.1: An example of two species showing Maximum, Mean, SD for abundance & abundance percentiles.

	Species <chr>	lower_95 <dbl>	lower_68 <dbl>	upper_95 <dbl>	upper_68 <dbl>	max_ab <dbl>	mean_ab <dbl>	sd_ab <dbl>
1	Ab	0	0	42.0	2.84	282	4	16
2	Jm+Bp	0	6	331.	141	1571	72	116

- Elevations that correspond to the abundance percentiles are identified. Columns namely "lower_95_el", "upper_95_el", "lower_68_el" and "upper_68_el" are the elevation corresponding to that abundance of 2.5th, 97.5th, 16th and 84th percentiles, respectively. Hence, these percentiles are used to try and identify the elevations corresponding to abundances that are approximately 2 and 1 standard deviations away from the mean. Elevations corresponding to abundance percentiles, the elevation of maximum abundance (max_el), mean abundance and standard deviation for abundance represented in Table 3.2.2 are the number of variables used for clustering the species

Table 3.2.2: An example of two species showing number of variables used in clustering species.

	Species <chr>	lower_ 95_el <dbl>	lower_ 68_el <dbl>	upper_ 95_el <dbl>	upper_ 68_el <dbl>	max_el <dbl>	mean_ab <dbl>	sd_ab <dbl>
1	Ab	0.620	0.620	2.03	2.07	1.24	4	16
2	Jm+Bp	0.270	0.270	2.53	2.64	1.93	72	116

.

- An example below clear emphasis on how the species distributions can be interpreted from Table 3.2.1, Table 3.2.2, and a plot Figure 3.2.1, which is representing characteristics of two species from these tables.

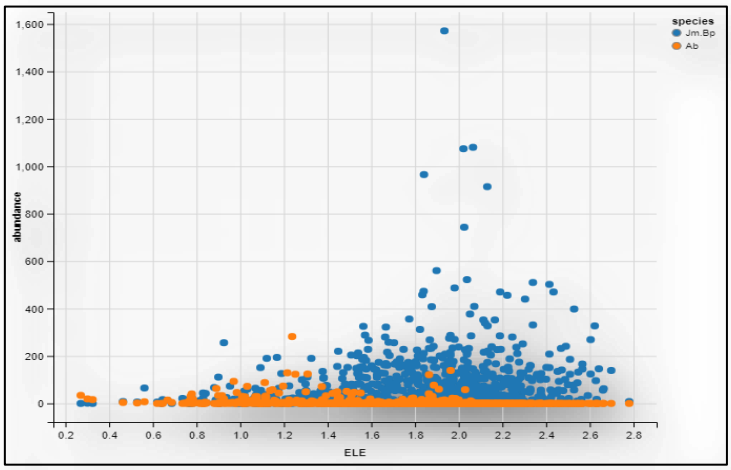

Figure 3.2.1: Species "Jm.Bp" & "Ab" distribution across elevation

Jm.Bp is a higher abundance species (max = 1571, mean = 72). It is highly abundant at middle elevations (~1.93) and has few or almost zero abundances at lower elevations (~0.27) and continues to have high abundance at higher elevations (~2.53).

Ab is a species with less abundance than Jm.Bp (max = 282, mean= 4). It has relatively low abundance at middle elevations (~1.24), and above middle elevations (~1.91), lower abundance towards higher elevations (~2.36) and zero abundance at lower elevations (~0.5).

STEP – 2: In this step, we cluster the data. The raw abundance data transformed into more useful data Table 3.2.2 which is suitable for clustering is used to cluster the species. Specific clustering approaches such as K-means clustering and Hierarchical clustering are used to group the species which will be discussed next in the report section. Species with similar variables (i.e., displays similar trends w.r.t elevation) will be grouped together.

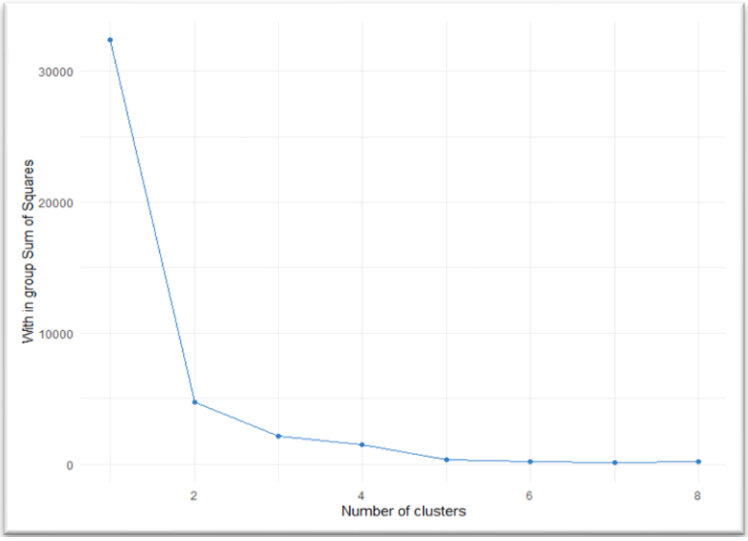

Figure 3.2.2: Scree Plot to determine number of clusters

The basic concept behind partitioning methods, such as k-means clustering, is to identify clusters in such a way as to reduce the total within-cluster sum of squares (WSS). The scree plot Figure 3.2.2 was used to evaluate the best k value that can be generated, i.e. number of clusters. The elbow at 4 clusters represents the most reliable and valid balance between reducing the number of clusters and decreasing the variance within each cluster.

3.2.1.a K- means Clustering

K-Means [8] [9] is one of the most common classification algorithms and is a common grouping tool that reduces the clustering error. K-means clustering is a common method used to divide a collection of data into k groups automatically. Whether the algorithm applies to the case or the dataset variables depends on the dimensions of this data set that we want to minimize. The aim is to generate groups of cases or variables that are highly similar within each group and that have minor similarities between groups [19].

In this analysis, K-means clustering is used as one of the clustering methods to group the species. The function Kmeans() [18] available in R was used to implement the K-means clustering algorithm and with parameter 'centres=K', to form the k number of clusters. Whereas k=4 is the number of clustering

groups assigned as illustrated in section 3.2.1 Clustering technique. Later the clustering data represented in Table 3.2.2 including a cluster number column allocated to each species is developed into a data frame. Species with identical variables showing equal variability in elevation will be grouped by the K-means Clustering method into the same cluster.

3.2.1.b Hierarchical Clustering

Hierarchical [9] classification is an algorithm grouping similar objects into groups known as clusters. The endpoint is a sequence of clusters, in which each cluster is different from the other, and in each cluster the objects, in general, are identical. The clusters generated in this procedure form a hierarchical structure of the tree kind. Hierarchical clustering [9] is a common method of clustering high-dimensional data that employs a cluster strategy, based on the difference between the observations of each pair. The algorithm begins with each observation in a different cluster and continues iteratively, joining at each step the two most "similar" clusters [12].

The dist() [18] function that is available in R was used to construct a distance matrix based on the Euclidean distance process. The hierarchical clustering algorithm was performed based on the Ward minimum variance method, using hclust() [18] which was found to be most appropriate as it creates a small number of clusters with relatively more species. Later the clustering data represented in Table 3.2.2 including a cluster number column allocated to each species is developed into a data frame. Species displaying the identical elevation pattern will be grouped by the hierarchical clustering process in the same cluster.

STEP – 3: Species are combined into groups by adding the counts of species assigned to the same group, which were determined by clustering methods (STEP-2), and thus forming a new dataset.

3.3 Method - III: Random Forest

“Random forests [6] are an ensemble learning method for classification, regression, and other tasks that function by building a multitude of decision trees at training time and outputting the class which is the classification mode or mean prediction (regression) of the individual trees” [6]. Random forests [6] are a method of combining different decision-making bodies that are trained in the various sections of the same training set to reduce the variance. It comes at the cost of a slight rise in bias and a certain lack of interpretability, but in general, the output in the final model improves significantly. More

specialized methods, which add up several trees to increase predictive efficiency are bagging, random forests and boosting. Bagging [6] is one way to boost the efficiency of a tree by measuring several trees and comparing the results from various samples. Bagging essentially follows the concepts of bootstrap. The steps are for b = 1…, B:

1. Sample n observations with replacement from the data.
2. Fit a tree $\hat{f}_k$ to the k[th] sample

In the regression, the setting combines the $\hat{f}_i$ using

$$\hat{f}(x) = \frac{1}{B}\sum_{k=1}^{b} \hat{f}_k(x)$$

Random forests [6] distinguish from this general algorithm in just one way: they use an algorithm to learn modified tree, which selects, separated by each applicant in the process of learning. A little variation on the above bagging technique is a random forest algorithm. The steps are for b =1..., B:

1. Sample n observations with replacement from the data.
2. Fit a tree $\hat{f}_k$ to the k[th] sample
3. At each step, in deciding on the optimal split, use only a random selection of the "m" available predictors. Where typically $m = \sqrt{P}$.

The Random Forest model described above is used to construct a standardized water level index predictive model using the randomForest() [6] function available in R. Random Forest function is used on the raw data to predict the SWLI value for each region of the United State using 10-fold cross-validation technique.

Here 10-fold cross-validation is carried out by diving the raw data into 10 individual parts. 9 parts of the training data are considered to train the model using random forest as the first phase of cross-validation, and the 10th individual part is considered to be the test data for predicting SWLI based on this training data. The cycle continues until every single part is considered as test data in the 10 splits.

3.3.1 Random Forest on Clustered Species

To explore the effect of reduced, twenty-three species count predictors variable clustering of species was implemented as a result new data set was created as explained in section 3.2.1 Clustering technique STEP – 3.

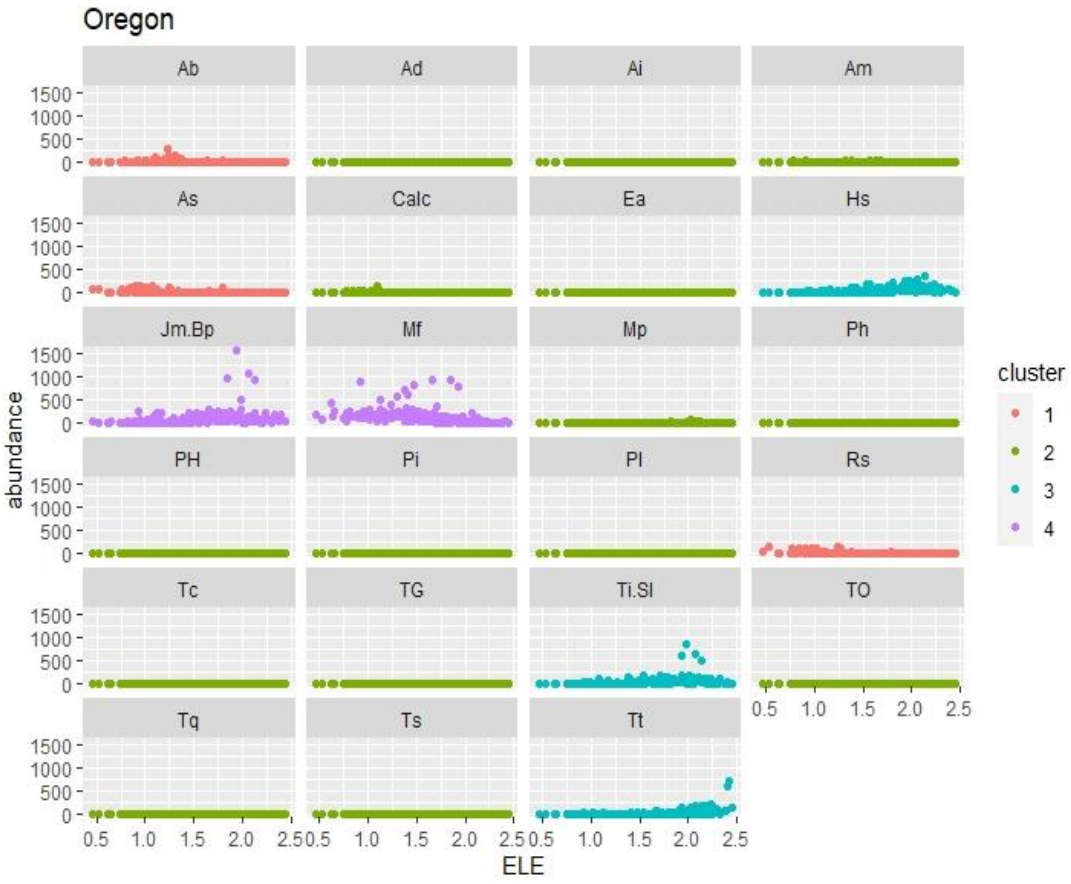

Figure 3.3.1: Clustered species exhibiting similar trend w.r.t elevation in Oregon region of United States.

Figure 3.3.1 was plotted using the tidyverse [20] package and the plot shows the species which are clustered in the Oregon region of United States by using Hierarchical Clustering method. The species that have similar variables (Table 3.2.2) are grouped together. The plot depicts that the species which are assigned to the same cluster exhibit similar trends w.r.t elevation and the colour distinguishes the clusters. For example, species in cluster 1 namely Ab, As and Rs show the same trend with respect to the elevation where their abundance is quite an in number from a lower elevation to middle elevations and zero abundance in higher elevation. Similarly, for cluster 2, cluster 3 and cluster 4 the species in their respective clusters are exhibiting similar trends with respect to elevation.

We add these clustered species counts together present in their respective clusters for each region of United States in the dataset using rowSums() [18] and as a result, four new columns in the dataset were created for the four clustered groups. For example, species counts in cluster 1 namely for Ab, As and Rs are summed together to form a new column “Cluster- 1”. Similarly, following this for other clusters we obtain a new clustered species dataset by combining all these species count data together for each cluster, giving us four species count clustered column variable namely cluster-1, cluster-2, cluster-3, and cluster-4 instead of twenty-three species count variable predictors.

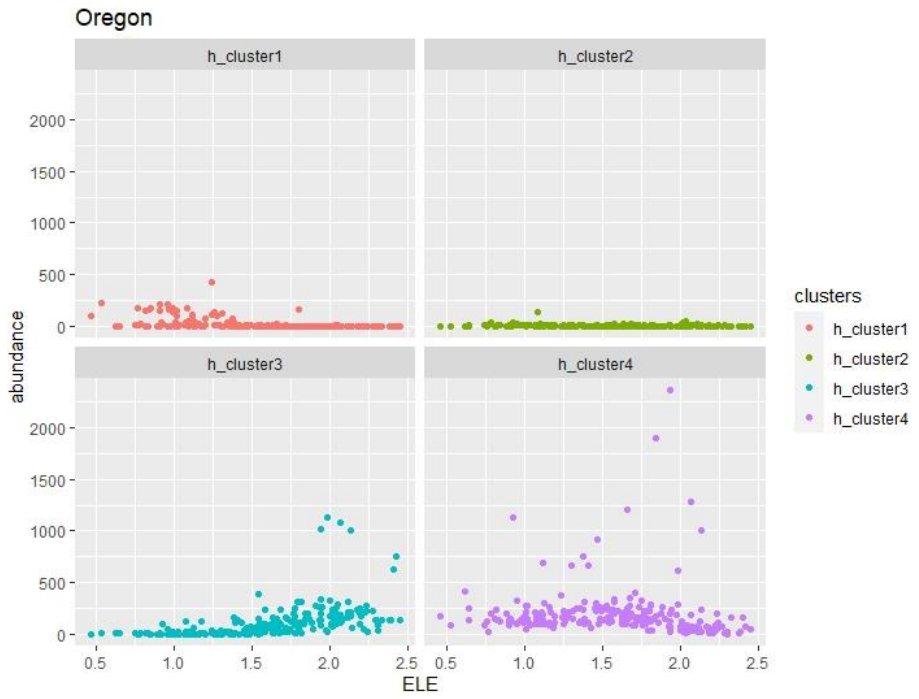

Figure 3.3.2: Summed clustered species exhibiting similar trend w.r.t elevation in Oregon region of United States.

Figure 3.3.2 illustrates the clustered group of species exhibiting the relationship with respect to elevation in the Oregon region of United States. Comparing Figure 3.3.1 and Figure 3.3.2 we could see that the clusters show a similar trend in elevation as of the individual species present in their respective groups. So, it is expected that the overall relationship with respect to elevation will be the same for the clustered species as it is for the individual species. Hence a new dataset was created by these clustered species columns as predictor variables. Random Forest function is applied to this new dataset created to predict SWLI and to explore the effect of reducing twenty-three species count predictors variables using a 10-fold cross-validation technique for each region of United States. The results and interpretation are illustrated in the later section of the report.

4 Results

This portion of the thesis work covers the findings obtained by implementing and exploring methods discussed in the above section 3 Methods. The empirical summary of observations for twenty-three species variables reduction approaches, i.e. PCA and clustering techniques – K means clustering and Hierarchical clustering, preceded by the findings of the Random Forest predictive model for raw data and new clustered species data.

4.1 Method – I: PCA Result

As stated in section 3.1 (Method - I: Implementing PCA), PCA was introduced as a result of the study of the impact of reducing the number of predictor variables for SWLI reconstruction. Here we shall explain the findings obtained through the implementation of PCA. As mentioned previously, by plotting a scree map, the number of components to be selected for the modelling stage is achieved. A scree plot is used for accessing components or variables that demonstrate the most data variation.

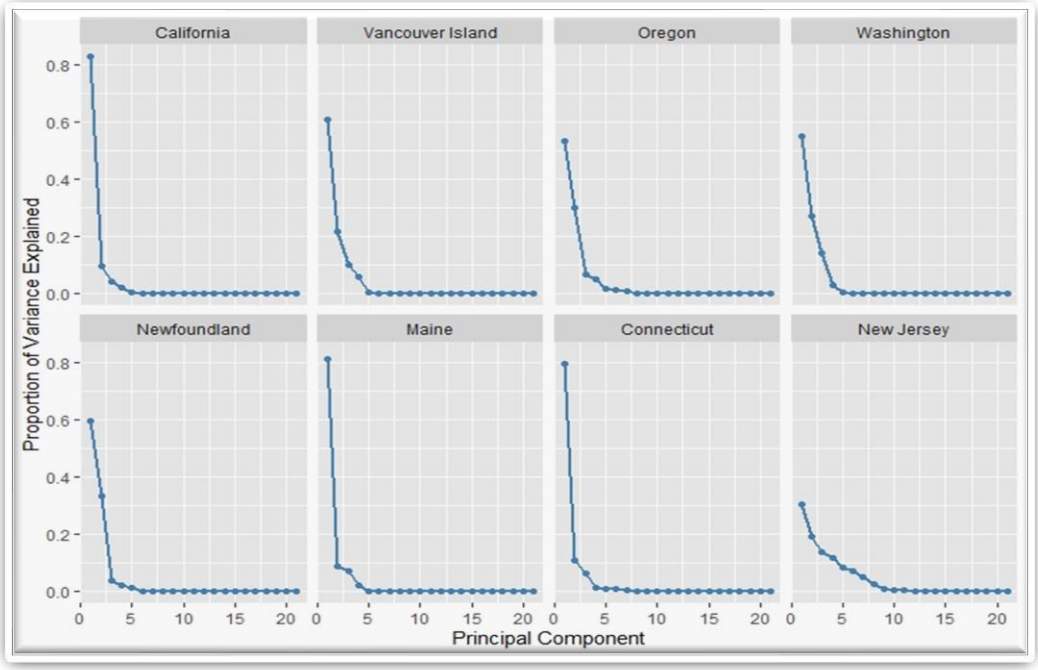

Figure 4.1.1: Scree Plot for each region of US to determine number of components to be used for modelling stage.

A scree plot demonstrates how much variance the data collect from each of the principal components. The y-axis is eigenvalues, which essentially stand for the amount of variation. The scree plot in Figure 4.1.1 above shows that about approximately 80% to 88% variation in the data set is explained by almost 2 to 3 components for each region of United States except for New Jersey. In scree plot Figure 4.1.1 we could see that for California, Maine and Connecticut first 2 components contain approximately 88% of the variance. Whereas for others first 3 components contain approximately 80% of the variance, while you need around 5 components to describe close to 100% of the variance for all regions. This plot shows that approximately 2 to 3 components result in variance close to 88%. Therefore, in this case, we would need about PC1 to PC2 or PC3 to retain 88% of the variance and to proceed for modelling stage.

4.2 Method – II: Clustering Results

The results obtained by implementing the clustering function from section 3.2.1 Clustering technique for "species grouping" are analysed by k-means and hierarchical clustering methods in this section.

4.2.1 K- means Clustering Result

k-means clustering method is applied to the group the species separately for each region initially and to see if the clustered species are similar across different regions.

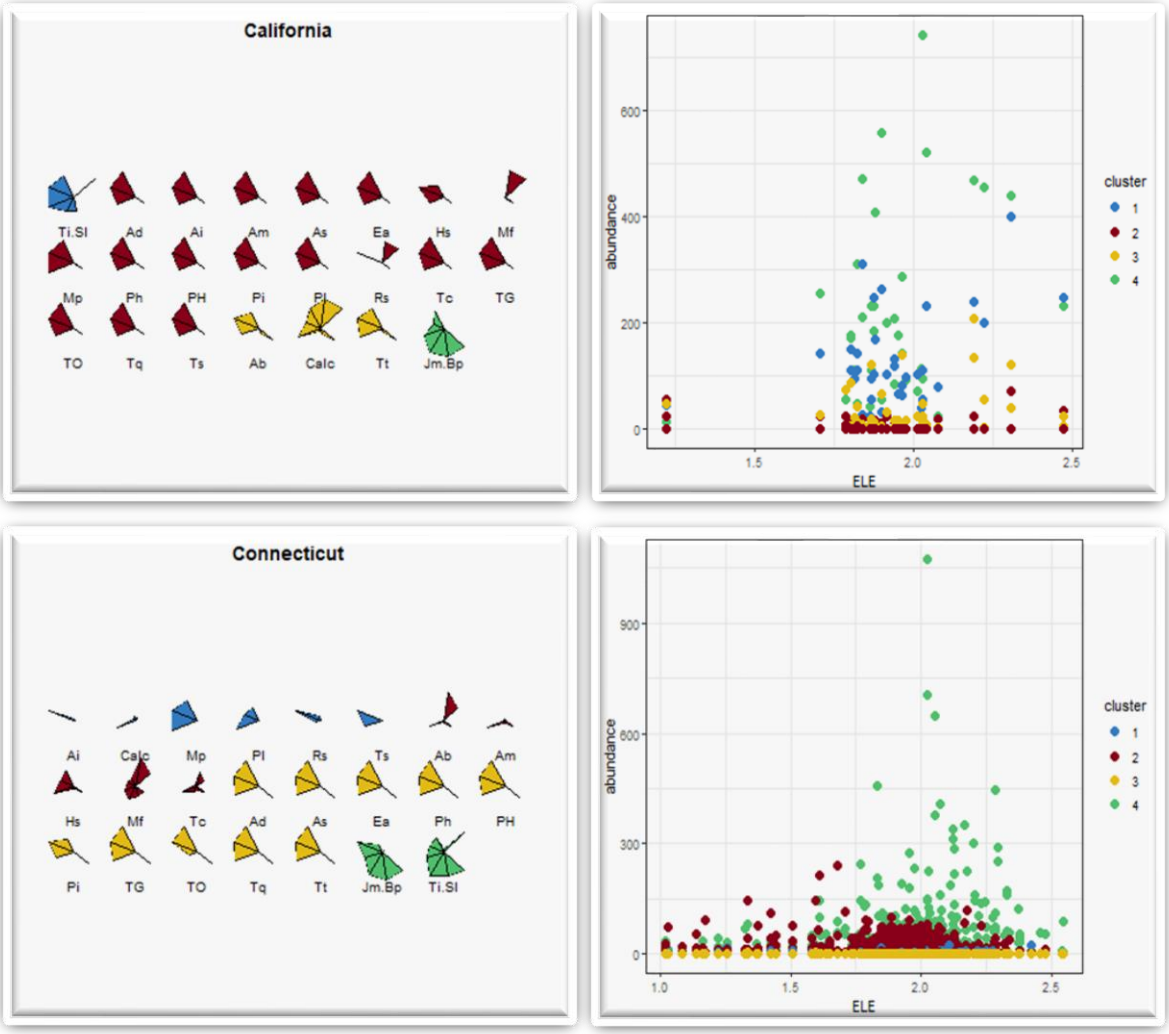

Figure 4.2.1: Species clustered by K-means Clustering Algorithm across California and Connecticut regions of United States

Figure 4.2.1 illustrates that in regions of Califonia and Connecticut –Jm.Bp, Ti.Sl is a higher abundance species (abundance ~10–800). It is most abundant above middle elevations (ELE ~1.75), has small abundance towards lower elevations (ELE ~1.5) and continues to have high abundance at higher elevations (ELE ~2.35).

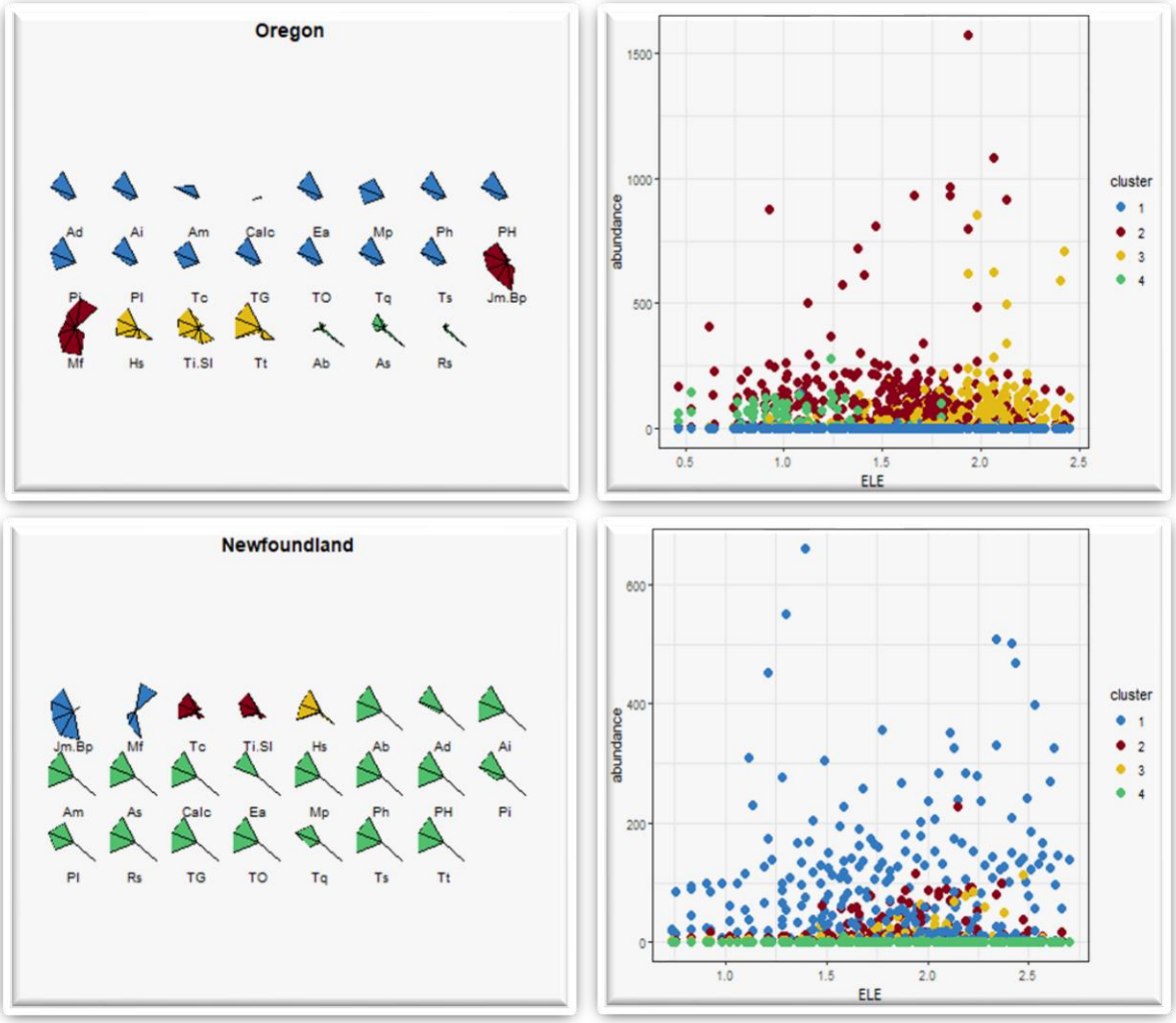

Figure 4.2.2: Species clustered by K-means Clustering Algorithm across Oregon and Newfound Land regions of United States

And Figure 4.2.2 demonstrates that in the region of Maine and Newfound Land – Jm.Bp, Mf is high abundance species (abundance ~10–1500). They have a fair amount abundance (abundance ~10–200) below middle elevations (ELE ~1.5), still relatively higher abundance above middle elevations (ELE ~1.75) and towards higher elevations (~2.36).

Whereas Hs is low abundance species (abundance ~10–150) in the region of Newfound Land and exists above middle elevations (ELE ~2.0). It has zero abundance at lower and higher elevations respectively. And rest other species have zero abundance across elevation.

While in the region of Oregon As, As, and Rs are the low abundance species (abundance ~10–200) and exist in small abundance towards lower elevations (ELE ~1.35). And species Jm.Bp, Ti.Sl and Mf continue to have higher abundance at higher elevations (ELE ~2.35).

4.2.2 Hierarchical Clustering Result

A hierarchical clustering method was used to group the species into four clusters with results summarised in the fan dendrogram Figure 4.2.3 below. R package ape [21] is used for plotting a fan type dendrogram type. And in our case, we have converted the hclust objects into phylo objects with the functions as.phylo. To illustrate, the species which have been assigned to the same groups are colour distinguished with respect to the other groups.

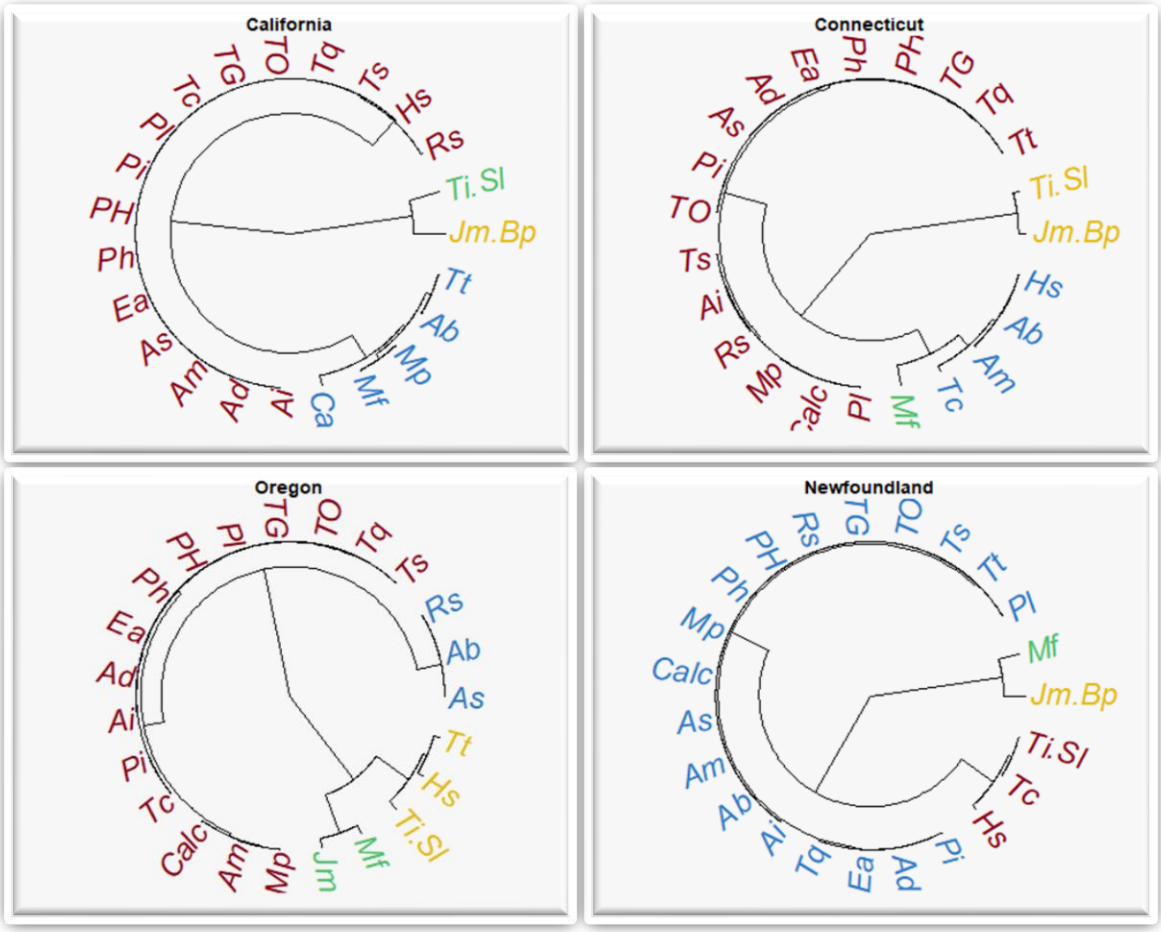

Figure 4.2.3: Species clustered by Hierarchical clustering algorithm across California, Connecticut, Oregon and Newfound Land regions of United States.

In the above Figure 4.2.3, at first Jm.Bp , Ti.Sl in regions of California and Connecticut are grouped in one single cluster, say cluster 1 or two separate individual clusters, say cluster 1 & 2, since they were the closest in exhibiting similar trends w.r.t elevation. Followed by species Mf and Hs in the regions of Oregon and Newfound Land are grouped with cluster 1 & 2 or forming an individual separate cluster, say cluster 3. The remaining species are combined to form another cluster, say cluster 4.

4.3 Random Forest Results

Random Forest function was applied to the raw data to predict the value of SWLI using a 10-fold cross-validation technique. Random forests are based on a basic concept that the average of the results of multiple predictors provides a better forecast than the best single predictor. We have trained and fit the model separately for each region of the US. And predicted the values for the test data of each region separately as explained in section 3.3 (Method - III: Random Forest).

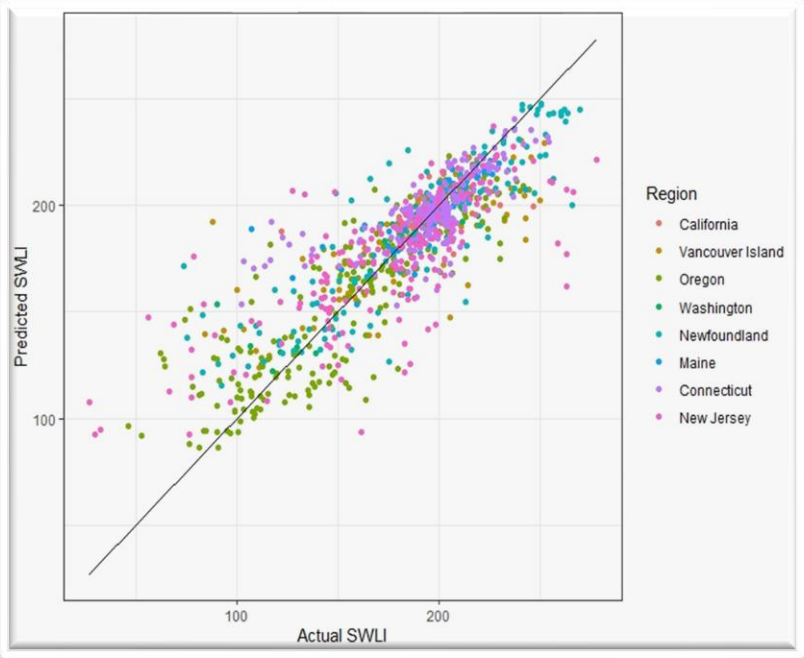

Figure 4.3.1: Actual versus Predicted Plot of Random Forest Regression.

Region	MAE
California	12.0899
Vancouver Island	18.54151
Oregon	16.33322
Washington	10.90807
Newfoundland	17.22616
Maine	11.33374
Connecticut	11.13790
New Jersey	22.62451

Table 4.3.1: MAE calculated for model fit by Random Forest for each region of United States.

The expected value is now compared with actual values in the test data and the model accuracy is evaluated. Figure 4.3.1 displays the predicted SWLI, points are almost close to this diagonal line (actual value) for all the regions, especially for the predicted value above 150 range of SWLI. Yet we must look at the mean absolute error (MAE) to more reliably evaluate our model. Table 4.3.1 illustrates the Mean Absolute Error calculated for each region of United States for model fit by Random Forest.

Figure 4.3.2 shows the Confidence interval plot plotted using rfinterval [22] R package indicating the precision of estimate made by Random Forest. Confidence intervals here include the point estimated for the SWLI sample with a margin of error around the point estimated. And 95% confident that the mean SWLI value will fall in the range of marginal error for each region.

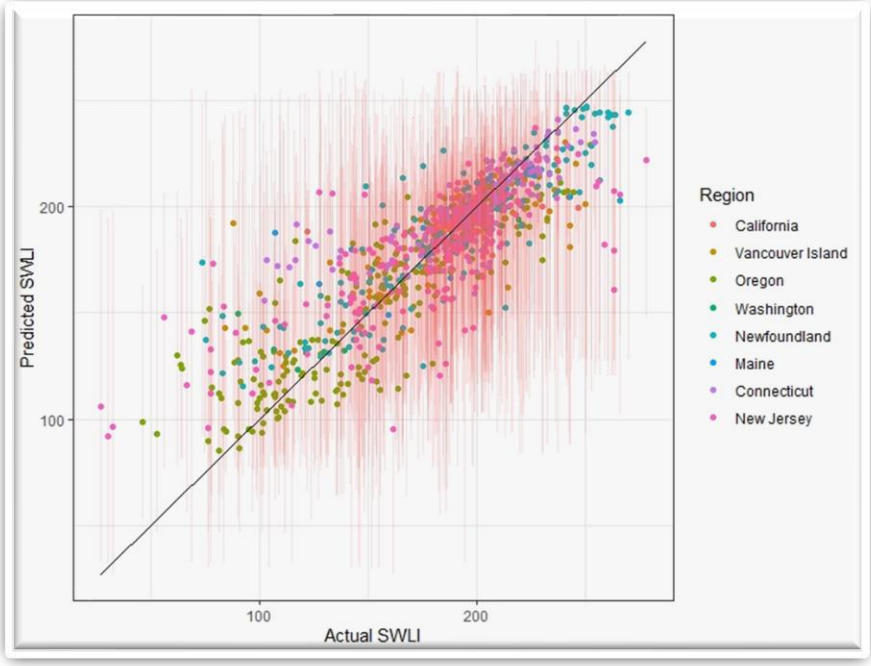

Figure 4.3.2: Confidence Interval of Predicted SWLI by Random Forest.

4.3.1 Random Forest on Clustered Species Result

As a result of applying Random Forest function on new dataset created for clustered species to predict the value of SWLI the below Table 4.3.2 demonstrates the Mean Absolute Error calculated when Random forest() is applied to clustered species data (new dataset) for both K means and Hierarchical clustering method and the raw (non-clustered) data across each region of United States. Also, the percentage of time the true SWLI falls within the confidence interval for each region of United States.

Table 4.3.2: MAE calculated for model fit by Random Forest for each region of U.S on clustered and non-clustered data and Percentage of time the True SWLI falls within CI for each region.

	Region	MAE_ Not_clustered	MAE_Kmeans	MAE_Hierarchical	Percentage_ Interval
1	California	12.01552	15.54025	13.51563	82.85714
2	Vancouver Island	18.60045	23.21645	22.05754	94.18605
3	Oregon	16.25845	19.19757	19.12693	95.37815
4	Washington	10.71177	12.49946	12.55693	94.11765
5	Newfoundland	17.65184	35.03623	18.88109	95.52239
6	Maine	11.62198	11.27045	11.37135	91.30435
7	Connecticut	11.14717	15.15358	14.71407	93.30709
8	New Jersey	22.56957	30.25337	31.00054	94.28571

5 Discussion

The analysis and findings carried out so far allow us to draw conclusions and explanations. The results obtained from different methods to lower the dimensional groups to predict elevation like PCA and clustering methods – K-means and Hierarchal clustering clearing state us the impacts of reducing the predictor variables. And we compared the result of the implementation of Random Forest, based on the abundance of species on the raw and clustered species dataset, to establish a predictive model for standardized water level index (SWLI) at different locations in the United States.

The results obtained in section 4.1 Method – I: PCA Result we ended up with 2 to 3 principal components, PCA may be one of the best ways to visualise the results, because we want to use the smallest possible number of principal components to explain most variation, by substantially reducing the number of variables using PCA. In the Proportion of variance plot Figure 4.1.1, the selected PCs were able to describe at least 80% of the variance for each region of United States except in the case of New Jersey. However, PCA is most suitable when variables have a linear relationship among them. The original features will be converted into principle components after PCA is introduced on the dataset. Principal Components cannot be read and perceived as the original feature in the dataset. While Principal Components aim to cover maximum variation among the features in the dataset, if we don't carefully select the number of Principal Components, some information might be missing compared to the original feature list. So instead, we considered other dimension reduction techniques like clustering.

We implemented the clustering technique as described in section 3.2.1 Clustering technique with the goal to divide the dataset into groups of similar characteristics, such that species in a group have similar variables as possible and are dissimilar to the species in another group. The two-clustering methods K-means and hierarchical clustering were used to interpret the outcomes. By using these two clustering methods there seems no big difference in the final results obtained in clustered species. However, by comparing the K-means clustering results from Figure 4.2.1 Figure 4.2.2 the differences in species exhibiting the trend with respect to elevation among the four clustered groups can be identified. In particular, from Figure 4.2.1 in the region of California from cluster 1 and cluster 4, and in the region of Connecticut from cluster 4 species namely Jm.Bp and Ti.Sl is higher abundance species exhibiting a similar trend with respect to elevation i.e they tend to exist in high abundance from middle elevation to higher elevations for these two regions. Whereas, from Figure 4.2.2 for the region of Oregon and Newfound Land from cluster 4 species namely As, As, and Rs are are the low abundance

species and tend to exist in low abundance in lower elevations and don't exist in higher elevations as compared to Jm.Bp, Ti.Sl and Mf are abundant in higher elevations for these two regions.

Also, in the case of the Hierarchical clustering method, the results in species clustering were similar to k-means clustering, from Figure 4.2.3 where Jm.Bp , Ti.Sl, Mf in regions of California, Connecticut, Oregon and Newfound Land are clustered to form individual or separate clusters whereas other species into separate ones. Analysing the clustering methods and based on species exhibiting trend with respect to elevation, the results indicate that certain species of foraminifera exist only in the higher elevation of the region of United States whereas few prefer the lower elevations across the coastal region of United States.

The current thesis used Random forests 3.3 Method - III: Random Forest as a predictive model in the reconstruction of the Standardised Water level index(SWLI). Random Forest provided an interpretable and accurate prediction model. As seen through the results from Figure 4.3.1 and Table 4.3.1 the Actual versus predicted plot and Mean Absolute error metrics, the model has strong predictive power with 10-fold cross-validation of the dataset for each region of United States. We can estimate the model fit by Random Forest for each region from Table 4.3.1 that mean absolute error calculated was least except for the region of New Jersey.

So far, the Random Forest was implemented in the raw dataset. In this thesis work, we also focused on the effect of lowering the dimensional groups by incorporating clustering methods for estimating the Standardised Water level Index and the random forest was applied using these clustered variables. As discussed in the previously to explore and analyse the impact we created a new dataset with the clustered species data as described in section 3.3.1 Random Forest on Clustered Species we interpreted from Figure 3.3.2 clusters showed a similar trend in elevation as of the individual species from Figure 3.3.1 present in their respective groups or clusters. So, it is anticipated that the overall relationship with respect to elevation for the clustered species will be equivalent to as it is for the individual species. From Table 4.3.2 we can compare and interpret that the Mean Absolute Error calculated for random Forest applied on a new clustered dataset containing a lowered number of predictor variables for the prediction of SWLI for each region, with the results of Mean Absolute Error on the raw dataset or non-clustered species dataset. Indeed, we found from the comparison of results Table 4.3.2, that the model has strong predictive power with 10-fold cross-validation of the dataset for each region of United States even with lower-dimensional predictor variables as well. Whereas the hierarchical clustering results were better compared to k-means clustering. Hence for future work with a larger dataset the lower the dimensional groups can be used in the reconstruction of SWLI.

6 Conclusion

In summary, we implemented Random Forest, based on the presence and abundance of certain species of foraminifera in salt marshes of the specified dataset, to establish a predictive model for standardized water level index (SWLI) at different locations in the United States. Analysed the impact of minimizing the number of species variables by "species clustering" using PCA and clustering techniques such as K- means clustering and hierarchical clustering. The relationship between the clustered species and tidal elevation was quantified, by observing the characteristics of the clustered species abundances with respect to elevation by a standardized calculation of elevation. We also compared the prediction model results that were developed with Random Forest for clustered and non-clustered data. By using the clustered species data in a prediction model resulted in better predictive SWLI as well. Resulting in using of lower-dimensional variables for the larger dataset in real practice for reconstruction of the Standardized Water level Index.

7 References

[1] I. Shennan, A. J. Long and B. P. Horton, "Introduction," in *Handbook of Sea-Level Research*, John Wiley & Sons, Ltd., 2015.

[2] O. van de Plassche, Sea-level research: A Manual for the Collection and Evaluation of Data., Norwich: GeoBooks, 1986.

[3] I. Shennan, "Handbook of sea-level research: framing research questions," in *Handbook of Sea-Level Research*, John Wiley & Sons, Ltd, 2015.

[4] R. Edwards and A. Wright, "Foraminifera," in *Handbook of Sea-Level Research*, John Wiley & Sons, Ltd., 2015.

[5] B. P. Horton and R. J. Edwards, "Quantifying Holocene Sea Level Change Using Intertidal Foraminifera: Lessons from the British Isles," *ScholarlyCommons,* vol. 40, p. 30, 2006.

[6] A. Liaw and M. Wiener, "Classification and regression by randomForest.," *R news,* vol. 2, no. 3, pp. 18-22, 2002.

[7] Y. Qu, G. Ostrouchov, N. Samatova and A. Geist, "Principal component analysis for dimension reduction in massive distributed data sets.," in *In Proceedings of IEEE International Conference on Data Mining (ICDM)*, 2002.

[8] T. Kanungo, D. M. Mount, N. S. Netanyahu, C. D. Piatko, R. Silverman and A. Y. Wu, "An efficient k-means clustering algorithm: Analysis and implementation.," *IEEE transactions on pattern analysis and machine intelligence,* vol. 24, no. 7, pp. 881-892, 2002.

[9] L. Rokach and O. and Maimon, "Clustering methods," in *Data mining and knowledge discovery handbook*, Springer US, 2005, pp. 321-352.

[10] A. Kemp, A. Wright, R. Edwards, R. Barnett, M. Brain, R. Kopp, N. Cahill, B. Horton, D. Charman, A. Hawkes, T. Hill and O. van de Plassche, "Late Holocene relative sea-level change in Newfoundland, Canada.," *Quaternary Science Reviews,* vol. 201, pp. 89-110, 2018.

[11] A. Kemp, B. Horton, C. Vane, C. Bernhardt, D. Corbett, S. Engelhart, S. Anisfeld, A. Parnell and N. Cahill, " Sea-level change during the last 2500 years in New Jersey, USA.," *Quaternary Science Reviews ,* vol. 81, pp. 90-104, 2013.

[12] A. H. A. Kemp, J. Donnelly, C. Vane, B. Horton, T. Hill, S. Anisfeld, A. Parnell and N. and Cahill, "Relative sea-level change in Connecticut (USA) during the last 2200 years.," *Earth and Planetary Science Letters,* vol. 428, p. 417-429, 2015.

[13] W. Chang, H. Wickham, M. Bostock and S. Décima, "Interactive Grammar of Graphics," 26 October 2019. [Online]. Available: https://cran.r-project.org/web/packages/ggvis/ggvis.pdf. [Accessed 02 07 2020].

[14] C. O. Wilke, “Package 'ggridges' - CRAN,” 12 01 2020. [Online]. Available: https://cran.r-project.org/web/packages/ggridges/ggridges.pdf. [Accessed 17 05 2020].

[15] S. Garnier, “viridis: Default Color Maps from 'matplotlib',” R package version 0.5.1, 2018.

[16] A. George, “Anomaly Detection based on Machine Learning: Dimensionality Reduction using PCA and Classification using SVM.,” *International Journal of Computer Applications,* vol. 47, 2012.

[17] A. Kassambara, “Principal Component Methods in R: Practical Guide,” Statistical tools for high-throughput data analysis, 2017.

[18] R. C. Team, “R: A Language and Environment for Statistical Computing,” R Foundation for Statistical Computing, Vienna, Austria, 2020.

[19] L. Morissette and S. Chartier, “The k-means clustering technique: General considerations and implementation in Mathematica,” *Tutorials in Quantitative Methods for Psychology,* vol. 9(1), pp. 15-24, 2013.

[20] Wickham et al., “Welcome to the tidyverse,” *Journal of Open Source Software,* vol. 4, no. 43, p. 1686, 2019.

[21] E. Paradis and K. Schliep, “ape 5.0: an environment for modern phylogenetics and evolutionary analyses in R.,” *Bioinformatics,* vol. 35, pp. 526-528, 2019.

[22] H. Zhang, “rfinterval: Predictive Inference for Random Forests,” R package version 1.0.0, 2019.

[23] F. Zolfaghari, H. Khosravi, A. Shahriyari, M. Jabbari and A. Abolhasani, “Hierarchical cluster analysis to identify the homogeneous desertification management units,” *PLOS,* 2019.

8 Appendix

We have used the R programming language for our project.