12-14-2016

Seeking Direct Coastal Erosion Associated Factors in the United States Shorelines

Le Chen

Le Chen

SEEKING DIRECT COASTAL EROSION ASSOCIATED FACTORS IN THE UNITED STATES SHORELINES

by

LE CHEN

Under the Direction of Jing Zhang, PhD

ABSTRACT

Sea-level rise is projected to have a wide range of effects on coastal environments, developments, and infrastructure. Based on the U.S. Geological Survey (USGS) Coastal Vulnerability Index (CVI) system data, we developed a two-stage model; firstly, the Bayesian Network (BN) is used to define relationship among driving forces; secondly, the logistic regression is used to evaluate direct association for direct factors related to Shoreline Erosion. Using this two-stage approach, increased sea-level (OR: 4.03[3.72,4.38]), higher Wave Height (OR: 0.56[0.54,0.61]), smaller Tidal Range (OR: 1.68[1.52,1.87]) and smaller Coastal Slope (OR: 0.45[0.44,0.49]) are directly associated with Shoreline Erosion in Atlantic Ocean; Geomorphology setting (OR: 9.35[6.33,14.18]) in high risk regions, such as beaches, is identified as direct association with Shoreline Erosion in Gulf of Mexico; Smaller tidal range (OR: 0.10[0.04,0.27] directly associated with Shoreline Erosion in Pacific Ocean. These direct factors were evaluated predictive ability with accuracy rates ≥ 0.59 and AUC ≥ 0.63.

INDEX WORDS: Bayesian Network, Direct Factors, Shoreline Erosion

SEEKING DIRECT COASTAL EROSION ASSOCIATED FACTORS IN THE UNITED STATES SHORELINES

by

LE CHEN

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

Master of Science

in the College of Arts and Sciences

Georgia State University

2016

SEEKING DIRECT COASTAL EROSION ASSOCIATED FACTORS IN THE UNITED STATES SHORELINES

by

LE CHEN

Committee Chair: Jing Zhang

Committee: Yi Jiang

Xin Qi

Electronic Version Approved:

Office of Graduate Studies

College of Arts and Sciences

Georgia State University

December 2016

ACKNOWLEDGEMENTS

I would like to appreciate a full blessing to my chair and advisor Dr. Jing Zhang, whom encouragement, guidance, and support from the initial to the final level enabled me to develop an understanding of applied statistics and theoretical mathematics.

I would like to thank my committee members Dr. Yi Jiang and Dr. Xin Qi, for their supports on my studies and thesis researches, insightful comments and suggestions. Special thanks go to my graduate director Dr. Gengsheng (Jeff) Qin and department chair Dr. Guantao Chen for most of my applied and theoretical foundations are built in the Department of Mathematics and Statistics.

During my two-year period as an IT Assistant for the Department of Mathematics and Statistics, many classmates, coworkers, and friends are helpful to support me. I have to acknowledge all persons in the Front Desk Office and Information Technology especially Mr. Tu Tran and Mr. Hubert White for their supports and assistances.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVATIONS

USGS – United States Geological Survey

BN – Bayesian Network

CVI – Coastal Vulnerability Index

MCMC – Markov Chain Monte Carol

NOS – National Ocean Service

WIS – Wave Information Studies

DCG – Directed Cyclic Graph

ABN – Additive Bayesian Network

ROC – Receive Operating Characteristic

AUC – Area under the Curve

CEIS – Coastal Erosion Information System

DAG – Directed Acyclic Graph

MLE – Maximum Likelihood Estimate

OR – Odd Ratio

AI – Artificial Intelligence

1 INTRODUCTION

1.1 Background and Purpose of the Study

1.1.1 Background

The Shoreline Erosion is a serious threat to waterfront property along the U.S. coastlines. The historical shoreline has been used to identify and evaluate potential shoreline changes [1]. The majority of those studies are designed to improve local coastal management and predicted potential response to the Sea-Level Rise. The Bayesian Network (BN) [2] [3] approach has been used in a variety of difference application, from studies of Artificial Intelligence (AI) to ecological system especially studies in bioinformatics. The objective of current study is to use a two-stage model, we developed to identify most important direct risk/protective factors contributing to Shoreline Erosion in the U.S. environmental science. Then, we evaluate the statistical predictive abilities of identified important direct factors.

1.1.2 Frequentist MLE and Bayesian methods

In parametric estimation methods: the frequentist Maximum Likelihood Estimation (MLE) and Bayesian Inference are often used. Both methods are asymptotically equivalent; but they are quite different in many aspects.

In MLE, we assume there is an unknown; but fixed parameter θ, and estimate θ and confidence interval. In here, θ is a point of an estimate value and it is not a random variable. However, in Bayesian Inference, we represent uncertainty about the unknown parameter θ, and use the probability to quantify these uncertainties. In here, θ is a set of probability distribution parameter and it is a random variable.

With a good prior information; the Bayesian procedure has a small-sample advantage. With either bad or no prior information; it may yield missing leading results for the data with a small sample size. The Bayesian method is often much computationally demanding and requiring various numerical methods such as Markov Chain Monte Carol (MCMC). The frequentist MLE is also computationally simpler; although it does not enjoy a small-sample advantage with a valuable sound prior information being wasted.

In summary, if the prior information is well-behaved and the sample size is enough larger. Both MLE and Bayesian prediction are in same converge; two approaches can yield similar results. The data and models grow complexity; however, two approaches can be diverging greatly. The MLE method has difficult to recover the expected values when the dataset is varying whereas the Bayesian Inference is closet the "true" values for all of scenarios.

1.1.3 Coastal Vulnerability Index (CVI) data

From historical to modern observations of a long-term shoreline change data is an ideal data set for the Bayesian statistical framework. Communicating information about effects of sea-level rise in terms of probability (Bayesian framework) may improve scientists' ability to address specific management questions regarding effects of Sea-Level Rise. The Coastal Vulnerability Index (CVI) [1] is developed to describe either physical processes or conditions at specific location along the U.S. coastlines. The six variables are defined as risk categories for the continental US coastlines are presented in Table 1.1, and these variables are used to develop and evaluate a BN to calculate probabilities of a long-term shoreline change [4], and tested over a two-year period in 2009 and 2010 [3] to predict vulnerability to Sea-Level Rise.

Table 1.1 Coastal Vulnerability Index Variables

Variable	Very Low 1	Low 2	Moderate 3	High 4	Very High 5
Geomorphology	Rocky, cliffs coasts, fjords	Medium cliffs, glacial, indented coasts	Low cliffs, Glacial drift, Alluvial plains	Cobble beaches, Estuary, Lagoon	Barrier beaches, sand beaches, salt march, mud flats, deltas, mangrove coral reefs
Coastal slope (%)	> 0.2	1.0 – 2.0	-1.0 – 1.0	-2.0 - -1.0	< -2.0
Relative sea level change (mm/year)	< 1.8	0.2 – 0.07	0.07 – 0.04	0.04 – 0.025	< 0.025
Erosion/accretion (m/year)	> 2.0	1.8 – 2.5	2.5 – 2.95	2.95 – 3.16	> 3.16
	Accretion		**Stable**	**Erosion**	
Mass wave height (m)	< 0.55	0.55 – 0.85	0.85 – 1.05	1.05 – 1.25	> 1.25
Mean tide range (m)	> 6.0	4.1 – 6.0	2.0 – 4.0	1.0 – 1.9	< 1.0

1.1.4 Purpose of the Study

Identifying direct associated risk factors contributing Shoreline Erosion by using a two-stage model we have developed. This identified information also can be used for a control of Shoreline Erosion, and for the selection of appropriate protection. It also helps to design a specific Shoreline Erosion control projects.

1.2 Expected Results

The identified direct/indirect and risk/protective factors for Shoreline Erosion are difference among Atlantic Ocean, Gulf of Mexico, and Pacific Ocean. We identified the best relationship among six variables: Geomorphology Setting (Geomorphic Risk), Shoreline Erosion (shoreline change rate), Coastal Slope (Coastal Slope), Sea-Level Change (Relative Sea-Level Rise), Mean Wave Height (Wave Height), and Mean Tide Range (Tidal Range) for Atlantic

Ocean, Gulf of Mexico, and Pacific Ocean. These identified variables' relationships are difference with relationships, which presented in a public report (Figure 1.1) [3]:

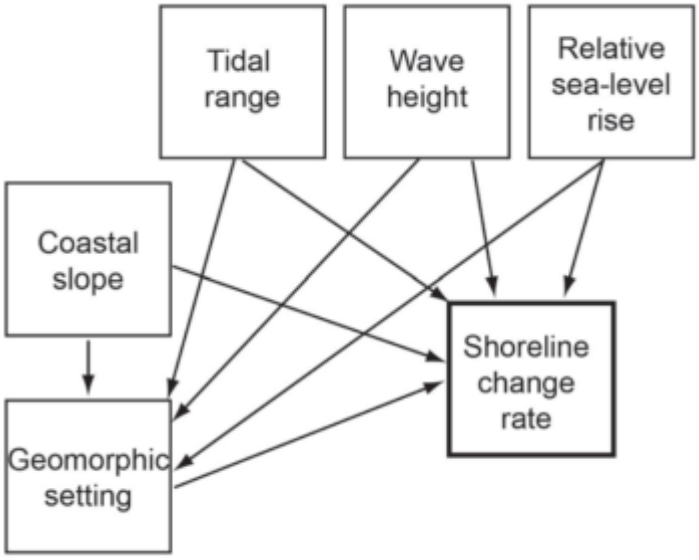


Figure 1.1 Diagram shown six variables' relationships

2 EXPERIMENT

2.1 Data

The project of coastal shorelines changes due to Sea-Level Rise collects data from rank rating system from Atlantic Ocean, Gulf of Mexico, and Pacific Ocean in the United States of America. The rating system is using the Coastal Vulnerability Index (CVI), which provides insight into the relative potential of coastal change due to the Sea-Level Rise. CVI allows six physical variables including Geomorphology Setting, Coastal Slope, Sea-Level Rise, Shoreline Erosion, Mean Tidal Range and Mean Wave Height (Table 1) to be related in quantifiable manner that expressed the relative vulnerability of the coast to physical change due to the Sea-Level Rise [5].

2.1.1 Variables description and resources

Geomorphology Setting:

This variable expressed a relative erodibility of different landform type. The data was derived from the State geological map and the United States Geological Survey (USGS): 1:250,000 scale topographic maps. Geomorphology Settings 1, 2, 3, 4, and 5 represents very low, low, modern, high, and very high vulnerability, respectively. The variables also are descripted in a link [3] and Table 1.1.

Shoreline Erosion (Accretion Rates):

The decadal-to-centennial scale historical rates of shoreline change based on data completed by May and others (1983)[3] and Dolan and others (1985)[3] into the Coastal Erosion Information System (CEIS). The CEIS data is drawn from a wide variety of sources included published reports, historical shoreline-change maps, field surveys and aerial-photo analyses. In

this analysis, this variable is a response variable and assume to be influenced by other five variables (Table 1.1) in the data set.

Coastal Slope:

This variable is estimated from the National Geophysical Date Center and U.S. Navy topographic and bathymetric data extending approximately 50 km landward and seaward of a local shoreline. It is also a measurement of the gradient of the substrate, which a local geomorphology had been formed and influenced the development of coastal landforms in region [3].

Sea-Level Change (Rate of Relative sea-level rise):

It is estimated by fitting a linear trend to the National Ocean Service (NOS) at a long-term (50-100 years) tide gauge observations and interpolating alongshore between NOS stations.

Mean Tidal Range:

It is estimated from the NOS at tide gauges and interpolated alongshore between NOS stations.

Mean Wave Height:

It is estimated from the U.S. Army Corps of Engineers' Wave Information Studies (WIS) hindcast data [3, 6] and interpolated between WIS stations.

Data Resource:

The Sea-Level dataset drives from difference systems and is stored in an attribute table associated with a 1:2,000,000 shorelines at three-minute resolution, which each three minute (~5 km) section of shoreline; there are six variables are merged into an observation data. This data is from a public USGS website [5] .

2.1.2 Missing Data

To better prepare data for BN analysis, we need to take parts into an account of the data's missing fact. "Amelia" is an R-package which provide visualized data and help checking data missing. Visualized plots for the Atlantic Ocean, Gulf of Mexico, and Pacific Ocean are displayed in Appendix A, B, and C, respectively. Note: one missing value was found and removed in the Gulf of Mexico dataset.

2.1.3 Correlations

In all of three datasets, we likely to know how six variables are related to each other. The Pairwise Pearson correlations are calculated Table 2.1, 2.2, and 2.3, and Figure 2.2, 2.3, and 2.4 displayed Pairwise Pearson correlations among six variables for Atlantic Ocean, Gulf of Mexico, and Pacific Ocean respectively.

Atlantic Ocean: Table 2.1 and Figure 2.1 present pairwise correlations for Atlantic Ocean project. The correlation between Shoreline Erosion (higher is more risk, Table 1.1) and Wave height (higher is more risk) is -0.04 (p value = 6.95E-6), -0.22 (p value = 2.2E-16) with sea level change (smaller is more risk, Table 1.1), 0.06 (p value = 5.46E-11) with tidal range (smaller is more risk, Table 1.1), and 0.04 (p value 6.37E-6) with costal slope (smaller is more risk). Shoreline Erosion did not show correlation with Geomorphic risk setting (correlation = 0.009, and p value = 0.3342).

Table 2.1 Atlantic Ocean's Pairwise Pearson Correlation Table

	Wave Height	**Tidal**	**Slope**	**Erosion**	**Sea Level**	**Geomorphic Risk**
Wave Height	1.00	0.16	0.41	**-0.04**	-0.97	-0.36
Tidal	0.16	1.00	0.59	**0.60**	-0.54	-0.22
Slope	0.41	0.59	1.00	**0.41**	-0.57	-0.50
Erosion	-0.04	-0.06	0.41	**1.00**	-0.22	0.01
Sea Level	-0.96	-0.54	-0.56	**-0.22**	1.00	0.29
Geomorphic Risk	-0.36	-0.22	-0.50	**0.01**	0.29	1.00

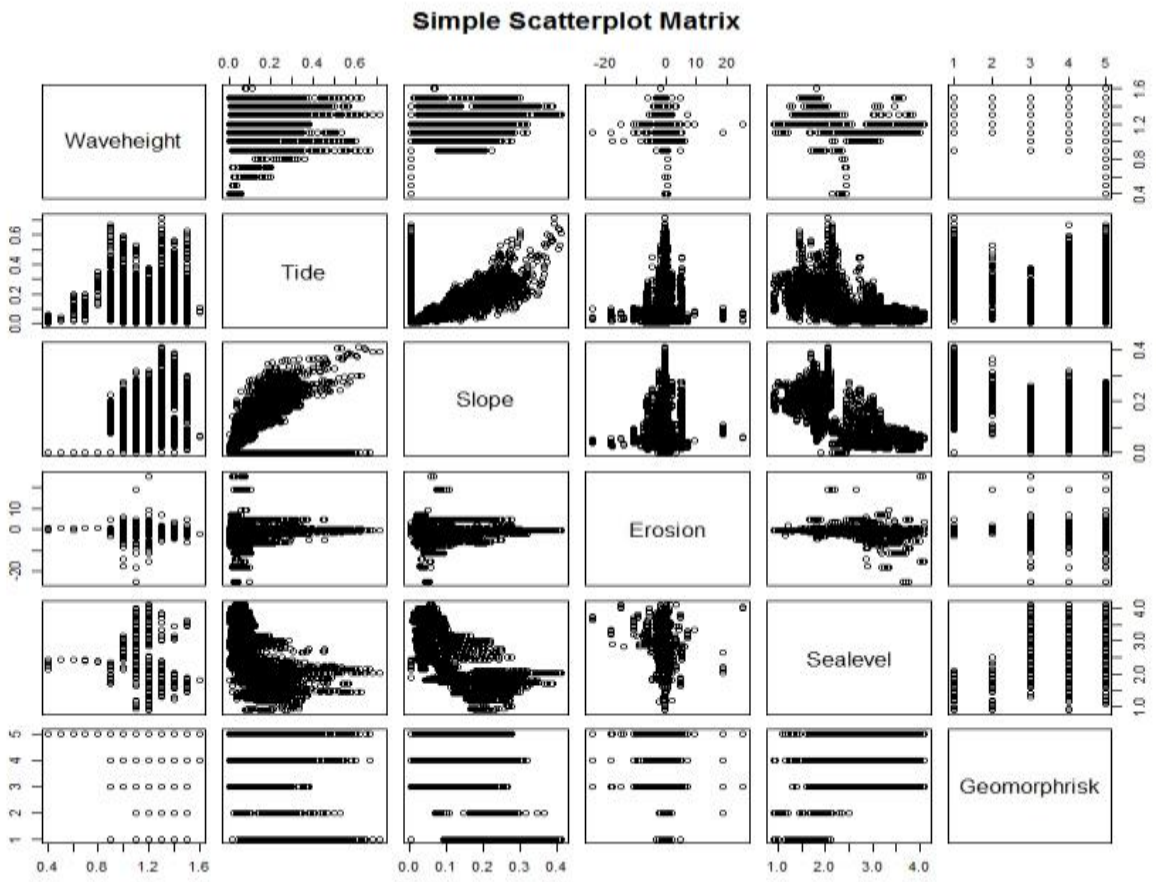


Figure 2.1 Atlantic's Ocean's Scatterplot

Gulf of Mexico: Table 2.2 and Figure 2.2 presents pairwise Pearson correlations for the Gulf of Mexico project. The correlations between Shoreline Erosion (higher is more risk, Table 1.1) and waive height (higher is more risk) is -0.22 (p value = 2.2E-16), -0.44 (p value = 2.2E-16) with sea level change (lower is more risk), -0.25 (p value = 2.2E-16) with Geomorphic risk setting (lower is more risk), and 0.26 (p value = 2.2E-16) with costal slope. Shoreline Erosion has no correlation with tidal range (0.04, p value = 0.1087).

Table 2.2 Gulf of Mexico's Pairwise Pearson Correlation Table

	Wave Height	**Tidal**	**Slope**	**Erosion**	**Sea Level**	**Geomorphic Risk**
Wave Height	1.00	-0.68	-0.30	**-0.22**	0.65	-0.08
Tidal	-0.68	1.00	0.02	**0.04**	-0.40	0.15
Slope	-0.30	0.02	1.00	**0.26**	-0.60	-0.18
Erosion	-0.22	0.04	0.26	**1.00**	-0.44	-0.25
Sea Level	0.65	-0.40	-0.60	**-0.44**	1.00	0.19
Geomorphic Risk	-0.08	0.15	-0.18	**-0.25**	0.19	1.00

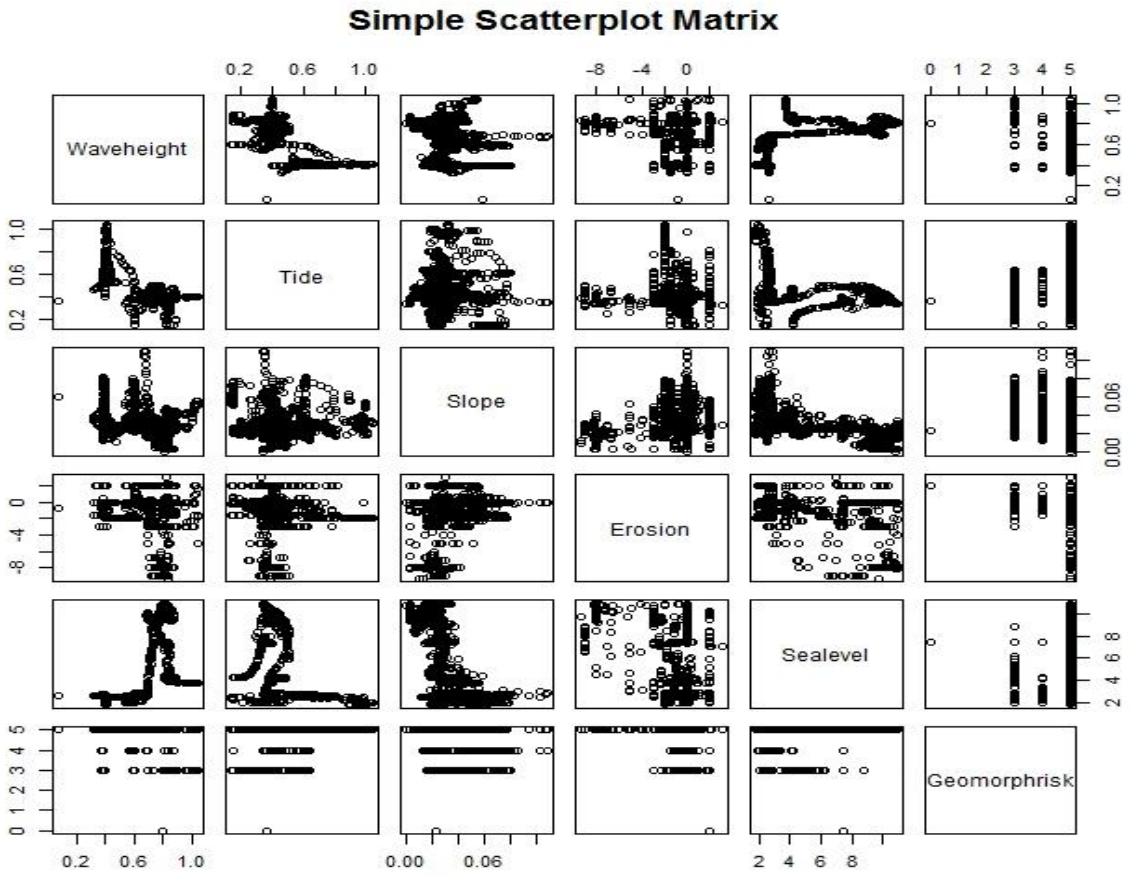


Figure 2.2 Gulf of Mexico's Scatterplot

Pacific Ocean: Table 2.3 and Figure 2.3 present Pearson correlation for Pacific Ocean project. No correlations between Shoreline Erosion with Wave height (correlation = 0, p value = 0.9737), tidal range (correlation = 0.03, p value = 0.2304), costal slope (correlation = 0, p value = 0.8826), and sea level change (correlation = 0, p value = 0.7754). Shoreline erosion show correlation with Geomorphic risk setting (correlation = 0.14, p value = 3.55E-8).

Table 2.3 Pacific Ocean's Pairwise Pearson Correlation Table

	Wave Height	**Tidal**	**Slope**	**Erosion**	**Sea Level**	**Geomorphic Risk**
Wave Height	1.00	0.68	-0.01	**0.00**	-0.68	0.22
Tidal	0.68	1.00	-0.09	**0.03**	-0.84	0.26
Slope	-0.01	-0.09	1.00	**0.00**	0.03	-0.30
Erosion	0.00	0.03	0.00	**1.00**	-0.01	0.14
Sea Level	-0.68	-0.84	0.03	**-0.01**	1.00	-0.19
Geomorphic Risk	0.22	0.26	-0.30	**0.14**	-0.19	1.00

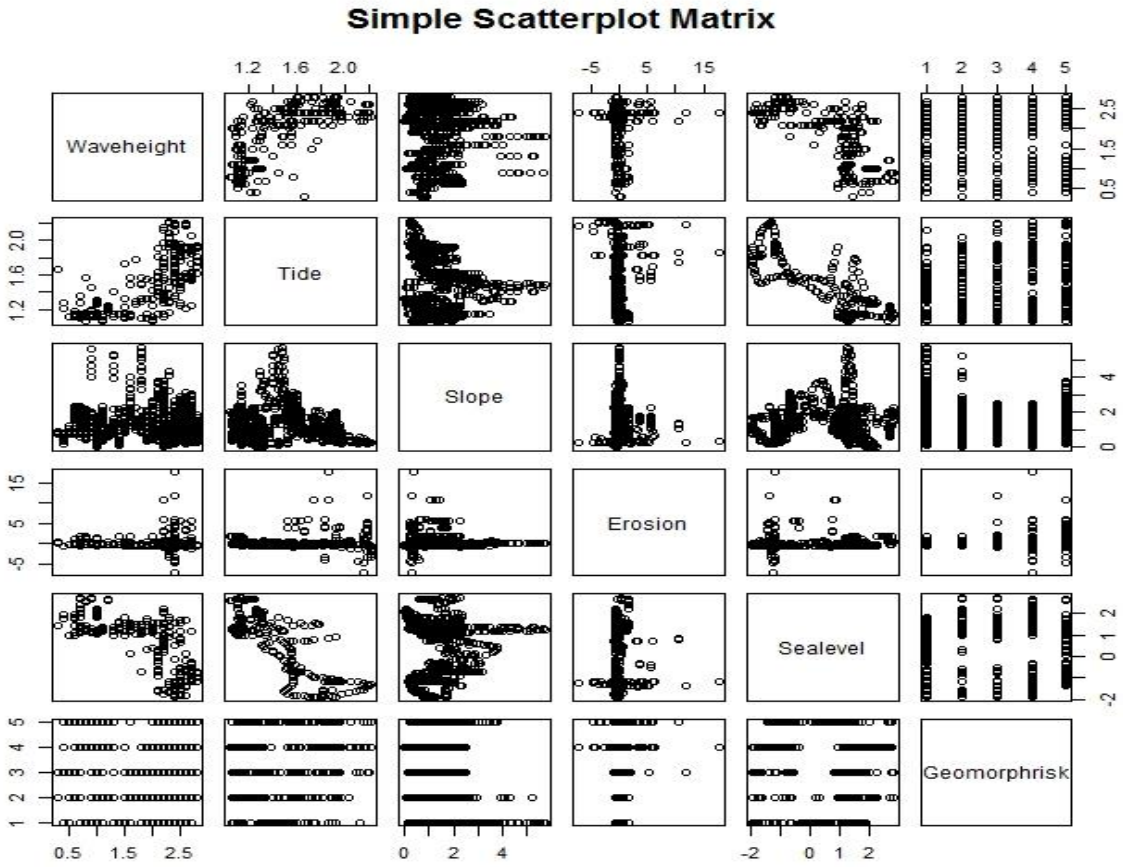


Figure 2.3 Pacific Ocean's Scatterplot

2.1.4 Distributions

There are two categories of random variables, which can be used in the BN models. The variables are checked normally firstly. Three figures: 2.4, 2.5, and 2.6 present distributions of six variables for Atlantic Ocean, Gulf of Mexico, and Pacific Ocean respectively. From QQ plots, all of six variables are not follow normal distributions in Atlantic Ocean, Gulf of Mexico, and Pacific Ocean. Those six variables are defined as binary traits based on Table 1.1.

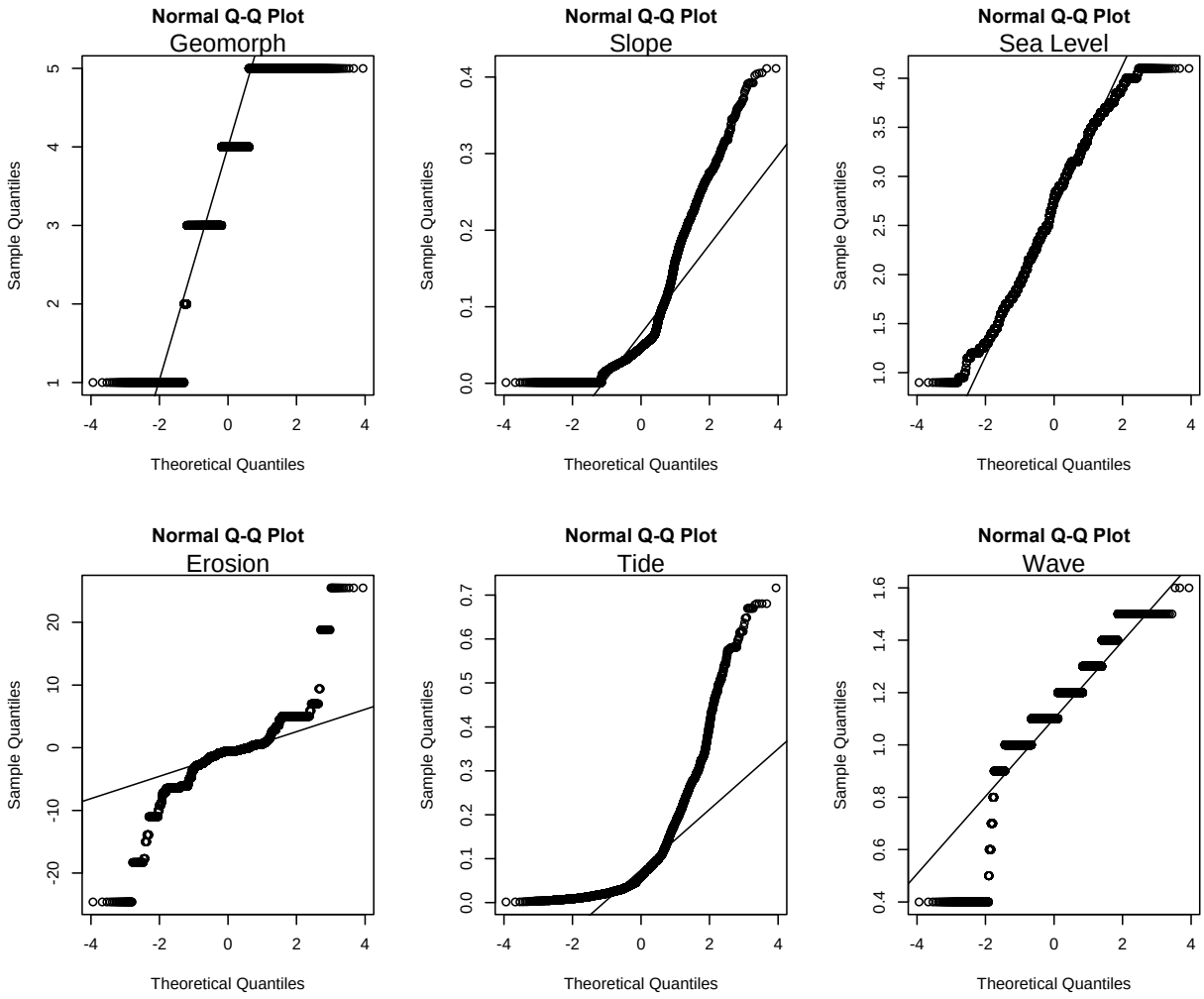


Figure 2.4 Atlantic Ocean's distributions

QQ plots of Geomorphology Setting, Coastal Slope, Sea-Level Change, Shoreline Erosion, Mean Tidal Range, and Mean Wave Height.

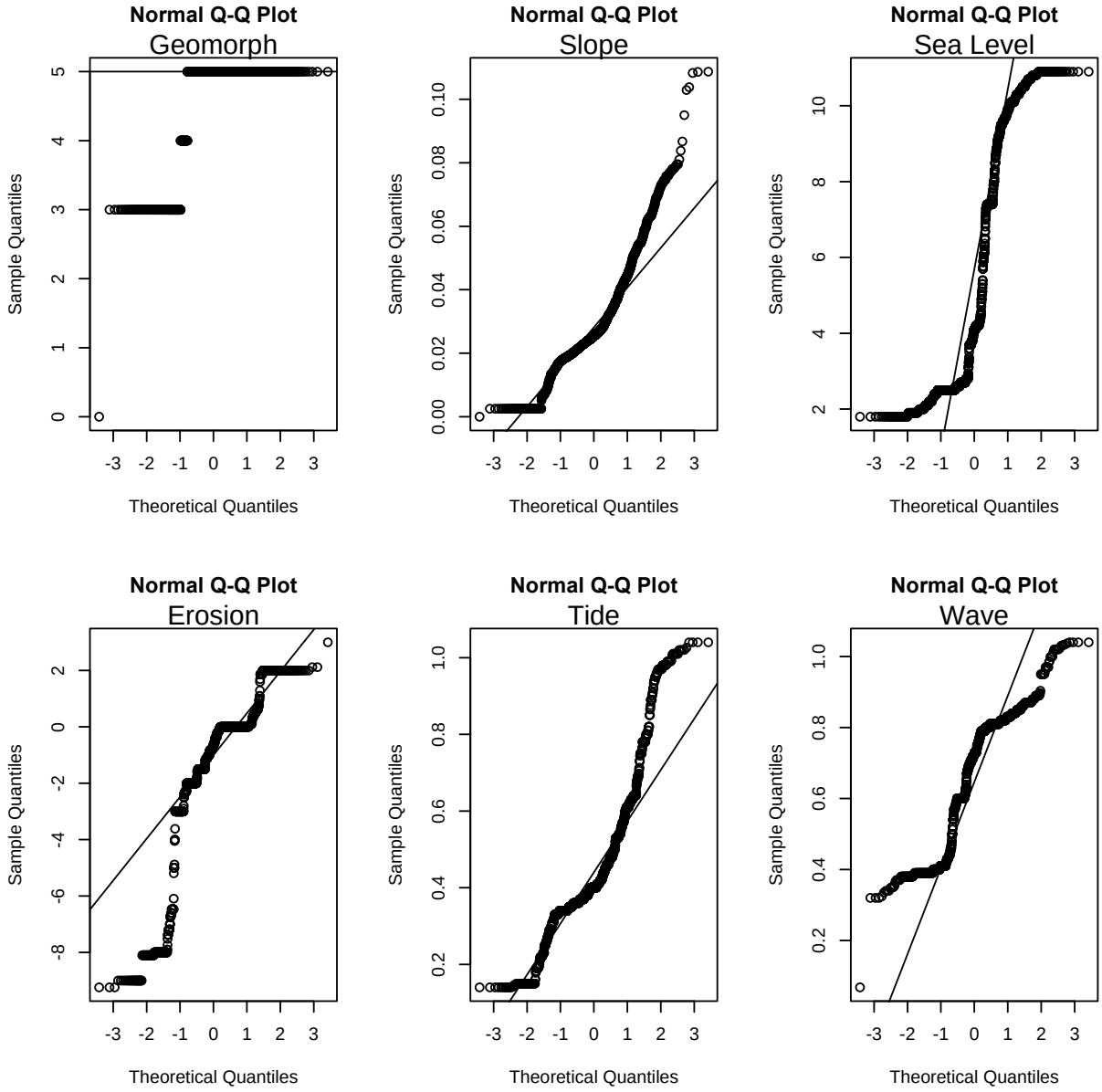


Figure 2.5 Gulf of Mexico's distributions

QQ plots of Geomorphology Setting, Coastal Slope, Sea-Level Change, Shoreline Erosion, Mean Tidal Range, and Mean Wave Height.

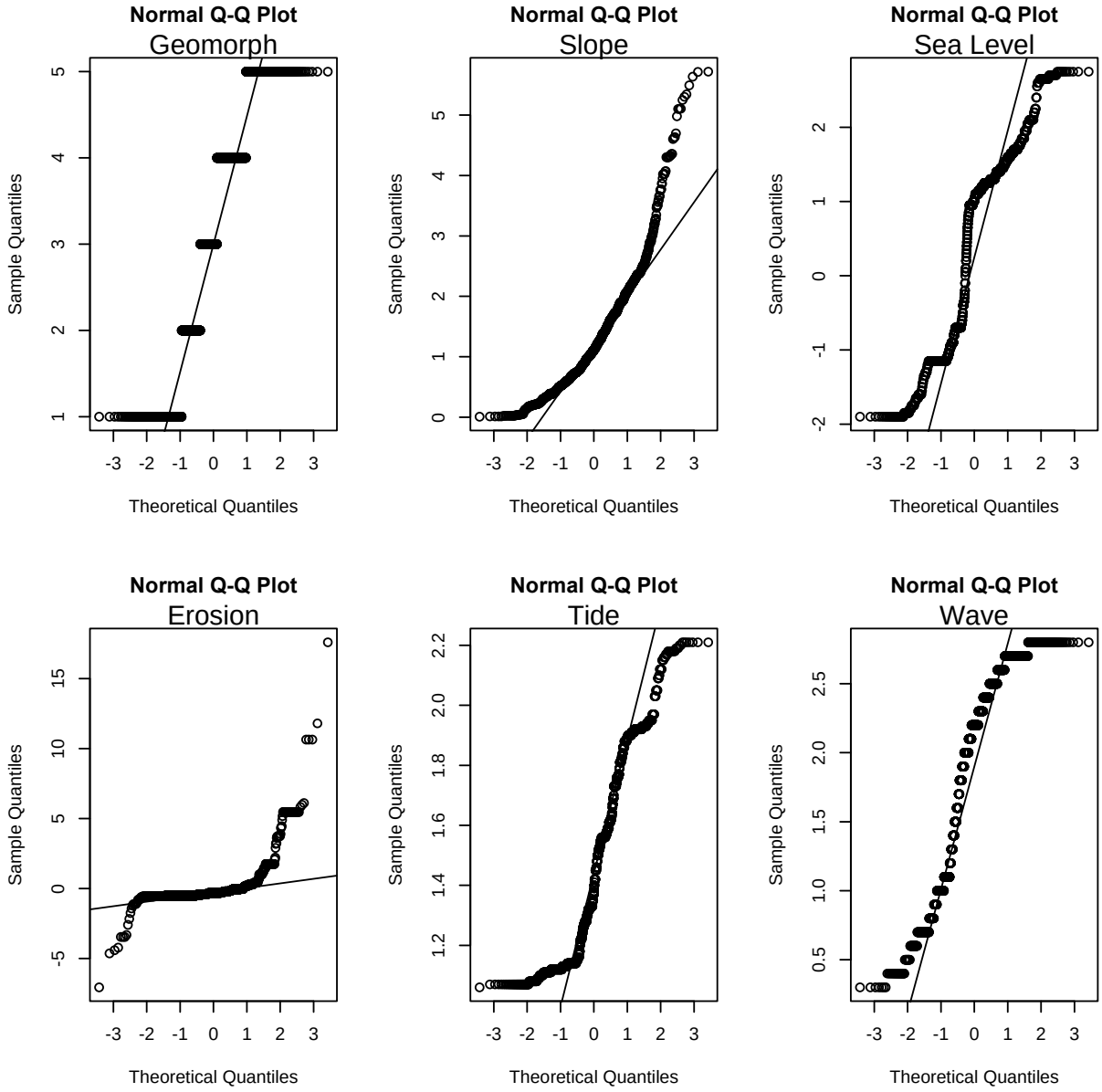


Figure 2.6 Pacific Ocean's distributions

QQ plots of Geomorphology Setting, Coastal Slope, Sea-Level Change, Shoreline Erosion, Mean Tidal Range, and Mean Wave Height.

2.2 Methods

2.2.1 General Bayesian Inference

$$p(R_i|O_j) = \frac{p(O_j|R_i) * p(R_i)}{p(O_j)}$$

With the formula above, it need to be expand with opposite side of prior and likelihood equations.

$$p(R_i|O_j) = \frac{p(O_j|R_i) * p(R_i)}{p(O_j|R_i) * p(R_i) + p(O_j|R_j') * p(R_j')}$$

2.2.2 General Bayesian Network (Stage 1)

$$L(\theta: Data) = \prod_v p(X_i[v], \dots, X_n[v]: \theta)$$

$$= \prod_i \prod_v p(X_i[v]|pa_i[v]: \theta) = \prod_i L(\theta : data)$$

In this study, the multivariate domains are available. The BN analysis is a style of probabilistic graphical models, which derived from empirical data: a directed acyclic graph (DAG) [7]. The two categories of random variables are used: first is multinomial data (discrete variable, binomial distribution) and second is multivariate normal data (continuous variable, normal distribution). The BN is a structure-learning algorithms along the conditional independence tests (constraint-based algorithms) and network scores (score-based algorithms), and it comprises three interrelated parts: 1) Parameter learning, 2) Network score, and 3) Structure learning [8]. Given a BN to fit whole random variables in the data set; a joint probability distribution is also eliminated. The parameter estimated process and scope, network scoring, and structure searching are well-documented and instructed by a simple example [9]. In here, we clean all of confused problems in BN:

A) Exhaustively searching with a larger number of variables; the number of BN structures is a super-exponential: $2^{\Theta(2^n)}$, where n is the number of variables. In order to reduce the complexity of computation; firstly, is to reduce the scope of the summation for both the marginal likelihood and the computation of feature probabilities. Secondly, each node (variable) as the score for some number of highest-scoring families are precomputed. Thirdly, reduce the cost of MCMC algorithm [9]. Those methods also can make BN available in small domains (typically 4-14 variables) [9].

B) Learning BN: an additive form of BN is used. The two approaches in structure discovery are used; firstly a heuristic local search approach [10] which is similar to a standard statistical multivariate regression. This regression is also used to identify high scoring and well-fitting models; other is to collapse DAGs over node ordering [11] and summarize results of BN model search. The benefit of searching across orders is to reduce search space from $\sim n!\, 2^{\frac{n}{2}}$ to $n!$ [12].

C) Arcs and direction: if two variables are correlated then it is likely that an edge between two variables will be appear in any high scoring model; the direction of the edge is naturally interpreted as the direction of causality. In BN, each DAG is formally a factorization of the joint probability distribution of nodes (random variables) since of likelihood equivalence; the presence of arcs between nodes is not a direction, which is a notable feature in DAG [11].

The "abn" of R-package is used for BN analysis [13].

2.2.3 Logistic Regression Model (Stage 2)

Logistic regression model provides an information which discuss about the relationship between response and exposure variables. From BN, identified direct variables with Shoreline Erosion are implemented into logistic regression model to determine the association between

Shoreline Erosion with the direct exposure variables. Logistic regression is also maximized the likelihood function of $p(Y|\theta)$, which definite to find a best parameter θ that maximizes how likely the observed data.

2.2.4 Assessing the predictive ability of direct effective variables

To assess a predictive ability of identified effective variables, we randomly split a whole dataset into two datasets: training data (60%) and testing data (40%). Firstly, the training dataset will be used to fit the model to estimate model parameters. These parameters also will be testing over the testing dataset to figure out how the model is working while predicting "shoreline erosion" on a new dataset. In a new dataset, we can compare predicted of Shoreline Erosion values with true Shoreline Erosion values and calculated accuracy rate, then plot a ROC (Receive Operating Characteristic) and calculated AUC (Area under the Curve).

2.2.5 R-code

We used R-package to analysis the datasets. All R codes for Atlantic Ocean, Gulf of Mexico, and Pacific Ocean projects can be found in Appendix D.

3 RESULTS

The basic study information is presented in Table 3.1: six continuous variables are presented in median (1^{st} and 3^{rd} quartiles). Based on variables of CVI risk rank-system, we had categorized variables into binary variables, the value in risk rank-system in 1 (very low), 2 (low), and 3 (moderate) at risk factor 1, and the risk rank-system in 4 (high) and 5 (very high) The risk factors' percentages are presented in Table 3.2.

Table 3.1 Basic Study Information

	Atlantic Ocean	Gulf of Mexico	Pacific Ocean
N	12288	1606	1635
Geomorphology	1(10%), 2(2%), 3(31%), 4(30%), 5(27%)	1(0%), 2(0%), 3(16%), 4(5%), 5(78%)	1(17%), 2(18%), 3(20%), 4(30%), 5(17%)
Coastal slope (%) **	0.4670 (0.0250, 0.1040)	.02500 (0.0197, 0.0366)	1.1130 (0.6810, 1.7360)
Slope Risk *	1(10%), 2(23%), 3(24%), 4(18%), 5(24%)	1(17%), 2(10%), 3(26%), 4(22%), 5(24%)	1(21%), 2(21%), 3(19%), 4(18%), 5(21%)
Relative Sea Level Change (mm/yr)	2.7500(2.1500, 3.1500)	4.1000(2.5000, 8.9000)	1.0000 (-0.9000, 1.4000)
Sea Level Change Risk *	1(11%), 2(33%), 3(16%), 4(16%), 5(24%)	1(1%), 2(8%), 3(0%), 4(46%), 5(45%)	1(23%), 2(27%), 3(44%), 4(5%), 5(0%)
Shoreline Erosion/accretion (m/yr)	-0.5000 (-2.2000, 0.2000)	-0.6485 (-2.000, 0.000)	-0.3000 (-0.5000, -0.0600)
Shoreline Erosion Risk *	1(10%), 2(2%), 3(48%), 4(13%), 5(27%)	1(1%), 2(8%), 3(46%), 4(15%), 5(30%)	1(3%), 2(5%), 3(90%), 4(1%), 5(1%)
Mean Tidal Range (m)	0.0596 (0.0278, 0.1207)	0.4000 (0.3500, 0.5300)	1.3700 (1.1400, 1.7400)
Tidal Risk*	1(0%), 2(1%), 3(20%), 4(29%), 5(50%)	1(0%), 2(0%), 3(1%), 4(0%), 5(99%)	1(0%), 2(0%), 3(3%), 4(97%), 5(0%)
Mean Wave Height (m)	1.1000 (1.0000, 1.2000)	0.7300 (0.4800, 0.8100)	2.2000 (1.3000, 2.5000)
Wave Height risk *	1(6%), 2(21%), 3(40%), 4(23%), 5(11%)	1(20%), 2(21%), 3(17%), 4(20%), 5(23%)	1(0%), 2(43%), 3(12%), 4(27%), 5(18%)

* category variables from 1 to 5; presented in frequencies
CVI risk rank-system: 1 = very low, 2 = low, 3 = moderate, 4 = high, and 5 = very high
** presented in median (1^{st} quantiles, 3^{rd} quantiles)

Table 3.2 Frequencies of risk factors

Variable	Atlantic Ocean	Gulf of Mexico	Pacific Ocean
Geomorphology (B_Geomo) *	57.31%	45.87%	83.74%
Coastal slope (B_Slope) (%)	42.69%	39.02%	46.39%
Relative sea level change (B_Sealevel) (mm/yr)	39.60%	6.00%	91.59%
Shoreline Erosion/accretion (B_Erosion) (m/yr)	37.89%	1.00%	45.26%
Mean Tidal Range (B_Tide) (m)	78.95%	96.50%	99.00%
Mean Wave Height (B_Wave) (m)	33.40%	45.20%	42.40%

* Based on CVI risk rank-system, 1 = very low, low, and moderate, 2 = high risk and very high risk
The percentages of risk factor 2 is presented.

The Pairwise Pearson correlations between Shoreline Erosion with Mean Wave Height, Sea-Level Change and Geomorphology Setting (from rocks to sand beaches) is a negative relationship. Shoreline Erosion with Mean Tidal Range and Coastal Slope is a positive relationship in Atlantic Ocean (Table 2.1 and Figure 2.2), Gulf of Mexico (Table 2.2 and Figure 2.3), and Pacific Ocean (Table 2.3 and Figure 2.4).

3.1 Atlantic Ocean

3.1.1 Bayesian Network (Stage 1)

The relationship among six variables in the Atlantic Ocean is presented in Figure 3.1. Coastal Slope, Sea-Level Change, Mean Tidal Range and Mean Wave Height showed a direct relationship with Shoreline Erosion. The Geomorphology Setting is not directly relationship with Shoreline Erosion.

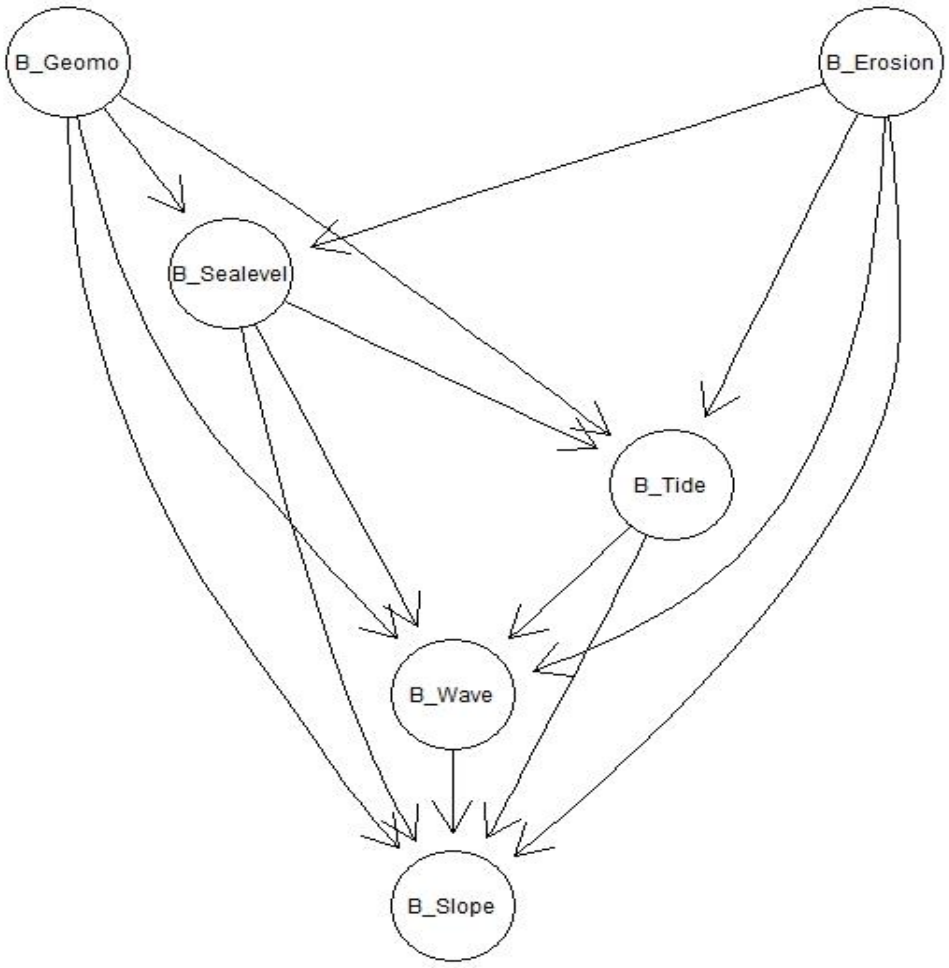


Figure 3.1 Atlantic Ocean BN variables' relationships

3.1.2 Logistic Regression Model (Stage 2)

The direct variables with Shoreline Erosion implemented into a final logistic regression model. In the analysis model, the response variable is a variable of Shoreline Erosion and binary variables of Coastal Slope, Sea-Level rise, Mean Tidal Range, and Mean Wave Height as explore variables. Table 3.3 shows association results with Shoreline Erosion for the Atlantic's coastlines. There are two protective factors: lower Coastal Slope and higher Wave Height are significantly associated with Shoreline Erosion with Odd Ratios of 0.45 [0.41, 0.49] and 0.56 [0.51, 0.61] respectively. There are two risk factors: Sea-Level Change and Mean Wave Height are significantly associated with Shoreline Erosion with Odd Ratios of 4.03[3.72, 4.40] and 1.68[1.52, 1.87] respectively.

Table 3.3 Atlantic Ocean's Direct Association Results

Variable	Estimate	Standard Error	Odd Ratios	95% CI (ORs)	Z-score	p-value
Coastal Slope	-0.803	0.045	0.449	(0.411, 0.489)	-18.144	<2e-16
Sea Level Change	1.39	0.042	4.03	(3.715, 4.375)	33.412	<2e-16
Mean Tidal Range	0.52	0.054	1.68	(1.516, 1.870)	9.759	<2e-16
Mean Wave Height	-0.579	0.047	0.56	(0.511, 0.614)	-12.364	<2e-16

* Based on Table separate each variable into two group, one is risk status in High and Very High, other is reference group

3.1.3 Assessing the predictive ability of direct effective variables

With the Table 3.3 above, the direct effective variables such as Coastal Slope, Sea Level Change, Mean Tidal Range, and Mean Wave Height are identified is also contributing to Shoreline Erosion. Next, it is need to understand predictive ability of these identified direct variables. The training dataset is used to fit the model, what it will be testing over the testing dataset. In the testing dataset, the predicted of Shoreline Erosion is compared with a true binary variable of shoreline erosion. The accuracy rate is 0.67 and the AUC is 0.71 (Figure 3.2).

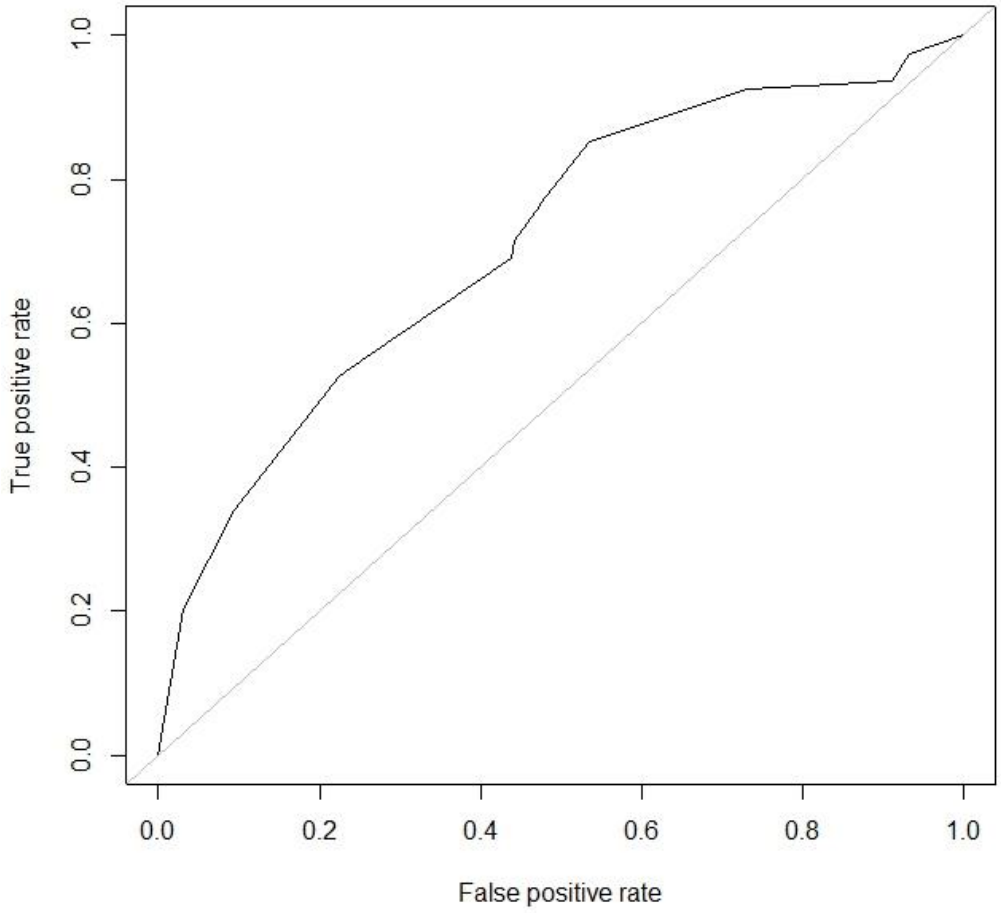


Figure 3.2 Atlantic Ocean ROC: AUC = 0.711

3.2 Gulf of Mexico

3.2.1 Bayesian Network (Stage 1)

The relationship among six variables in Gulf of Mexico project is presented in the Figure 3.3: it displayed that the outcome has direct relationship with Mean Tidal Range, Geomorphology setting, and Sea-Level Change showed direct relationship with Shoreline Erosion.

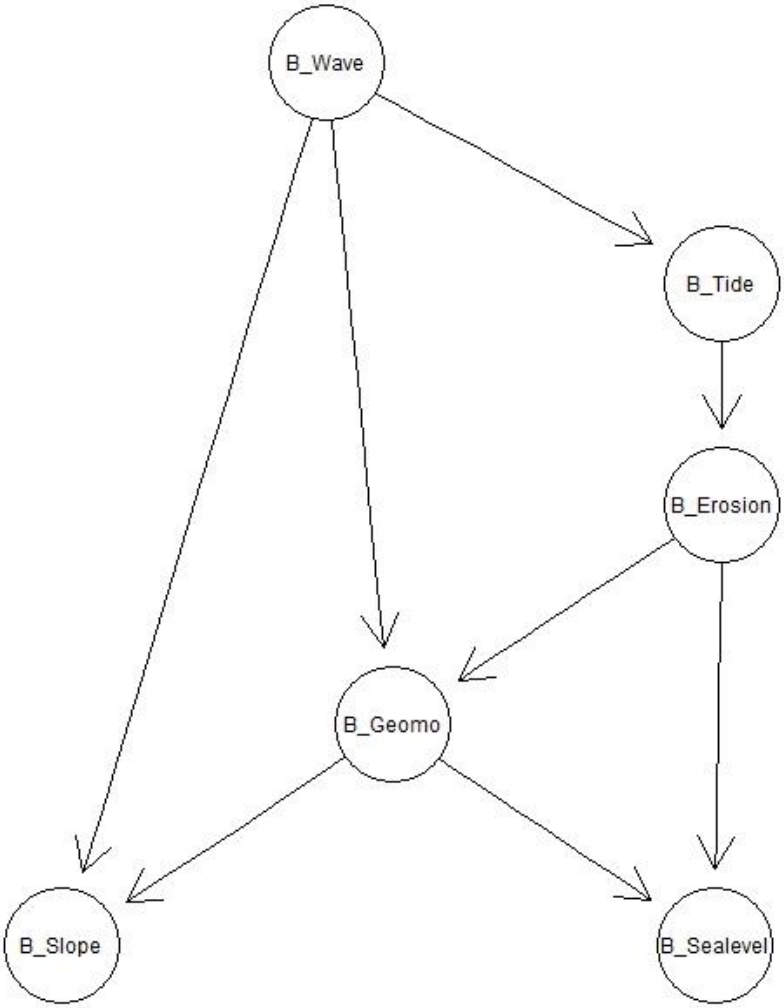


Figure 3.3 Gulf of Mexico BN variables' relationships

3.2.2 Logistic Regression Model (Stage 2)

The direct relationship variables with Shoreline Erosion: Mean Tidal Range, Sea-Level Change and Geomorphology Setting as independent variables and Shoreline Erosion as a dependent variable are implemented into a final regression model. The Table 3.4 showed direct association results with Shoreline Erosion. Mean Tidal Range and Sea-Level Change are not associated with Shoreline Erosion.

The Geomorphology Setting such as cobble, barrier beaches, estuary, lagoon, sand beaches, salt beaches, mud flats, deltas, mangrove, and coral reefs is significantly associated with Shoreline Erosion with ORs: 9.34[6.33, 14.18]. The associated results table is presented in Table 3.4:

Table 3.4 Gulf of Mexico's Direct Association Results

Variable	Estimate	Standard Error	Odd Ratios	95% CI (ORs)	Z-score	p-value
Geomorphology Setting	2.2351	0.2032	9.347	(6.334, 14.178)	10.997	<2E-16

3.2.3 Assessing the predictive ability of direct effective variables

Along the 3.2.2 section, the direct effective variable of Geomorphology Setting with Shoreline Erosion's significant association is identified. To accessing the predictive ability of the Geomorphology Setting, a whole data is spited into two datasets: training and testing. Firstly, the training dataset is used to fit a model, which will be tested over the testing dataset. Secondly, the testing dataset; the predicted of Shoreline Erosion is compared with a true binary variable of the Erosion. The accuracy rate is .59 and the AUC is 0.63 (Figure 3.4).

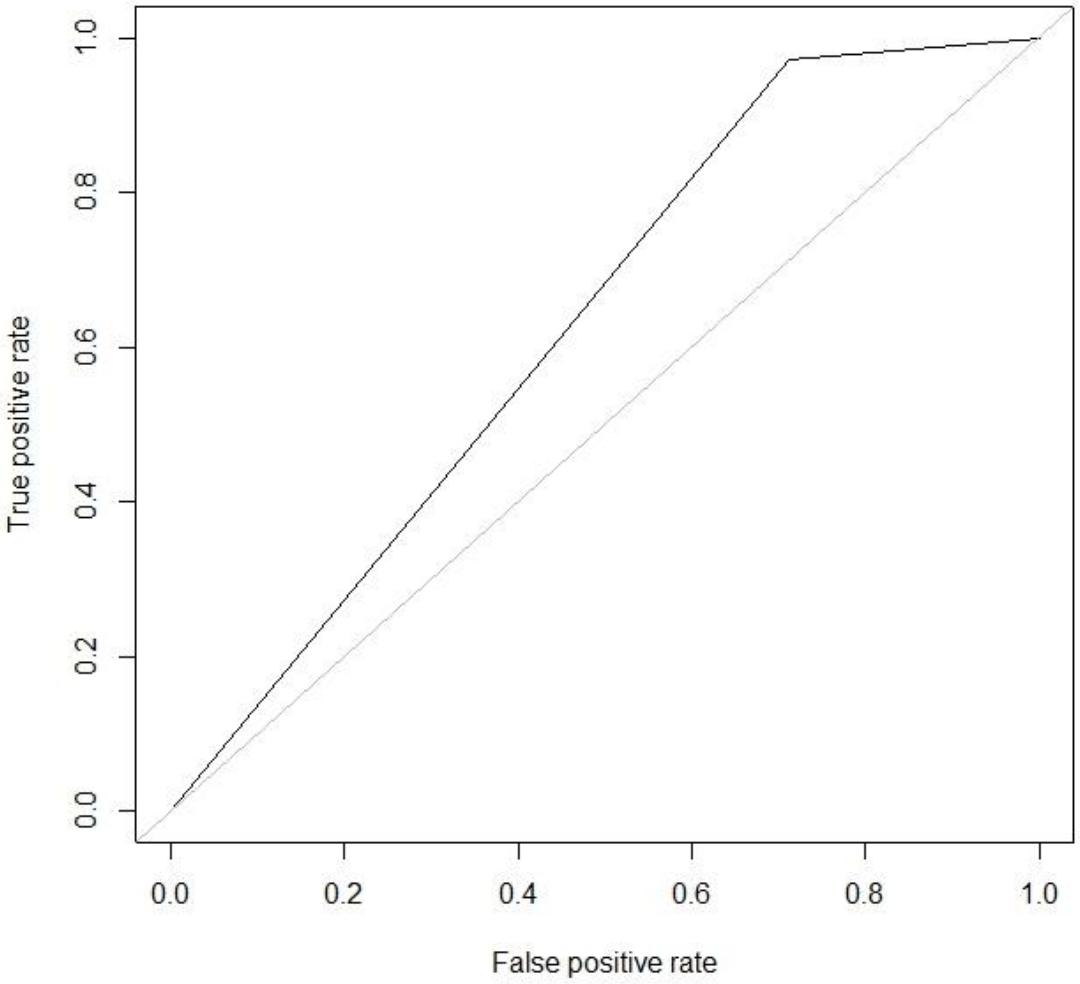


Figure 3.4 Gulf of Mexico ROC: AUC = 0.63

3.3 Pacific Ocean

3.3.1 Bayesian Network (Stage 1)

The relationship among six variables in the Pacific Ocean project is presented in the Figure 3.3.1: Geomorphology Setting, Mean Wave Height, Coastal Slope and Mean Tidal Range showed direct relationship with Shoreline Erosion. Sea-Level Change, is indirectly relationship with Shoreline Erosion.

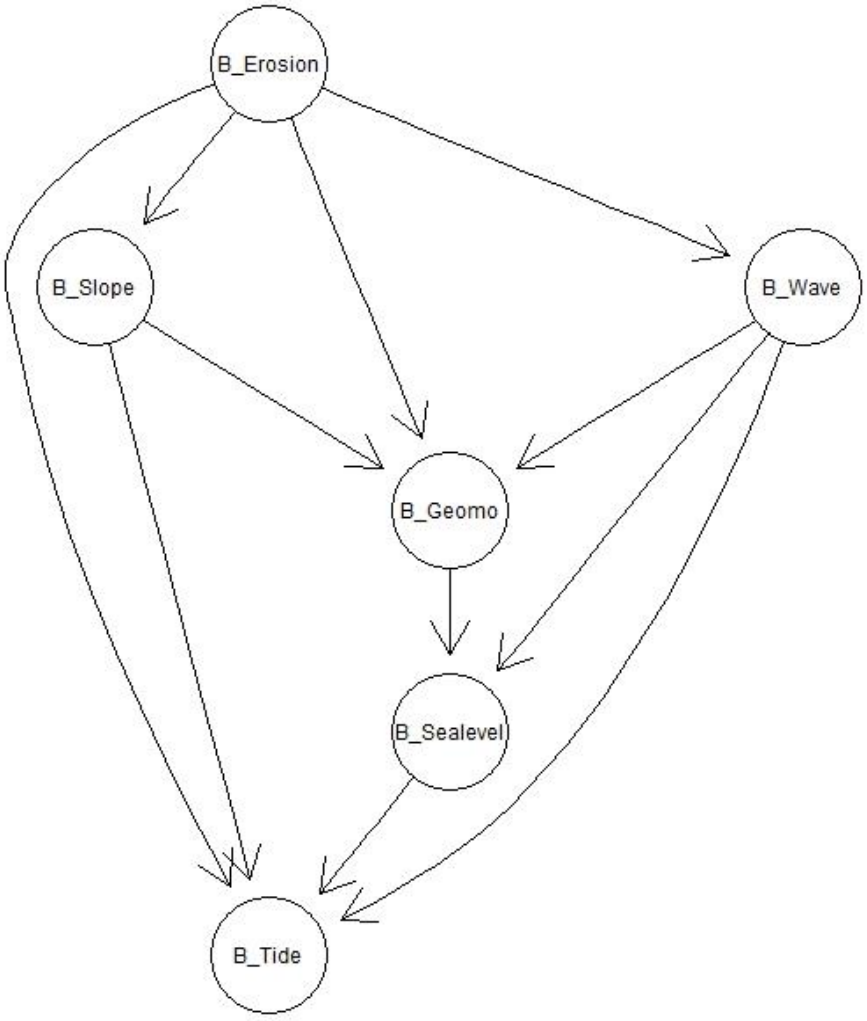


Figure 3.5 Pacific Ocean BN variables' relationships

3.3.2 Logistic Regression Model (Stage 2)

The direct relationship variables with Shoreline Erosion are: Coastal Slope, Mean Wave Height, Mean Tidal Range, and Geomorphology Setting since Shoreline Erosion is a dependent variable. Both direct relationship and dependent variable are implemented into a final regression model. The Table 3.5 showed variables' direct association results with Shoreline Erosion.

In results, three variables: Coastal Slope, Mean Wave Height, and Geomorphology Setting are indirectly association factors with Shoreline Erosion. The smaller Mean Tidal Range showed significantly protected association factor with Shoreline Erosion at Odd Ratios at 0.10 [.0036, .2734]. The Table 3.5 is displayed below:

Table 3.5 Pacific Ocean's Direct Association Results

Variable	Estimate	Standard Error	Odd Ratios	95% CI (ORs)	Z-score	p-value
Mean Tidal Range	-2.3453	0.5503	0.096	(.0036, .2734)	-4.262	2.03E-05

3.3.3 Assessing the predictive ability of direct effective variables

From the Table 3.5, only one directly effective variable, Mean Tidal Range; which associated with Shoreline Erosion is identified. To accessing the predictive ability of the variable, a whole data is split into two datasets: training and testing. The training dataset is used to fit the model, which it will be testing over the testing dataset. In the testing dataset, the predicted of shoreline erosion is compared with a true binary variable of the erosion. The accuracy rate is 0.98 and the AUC is 0.91.

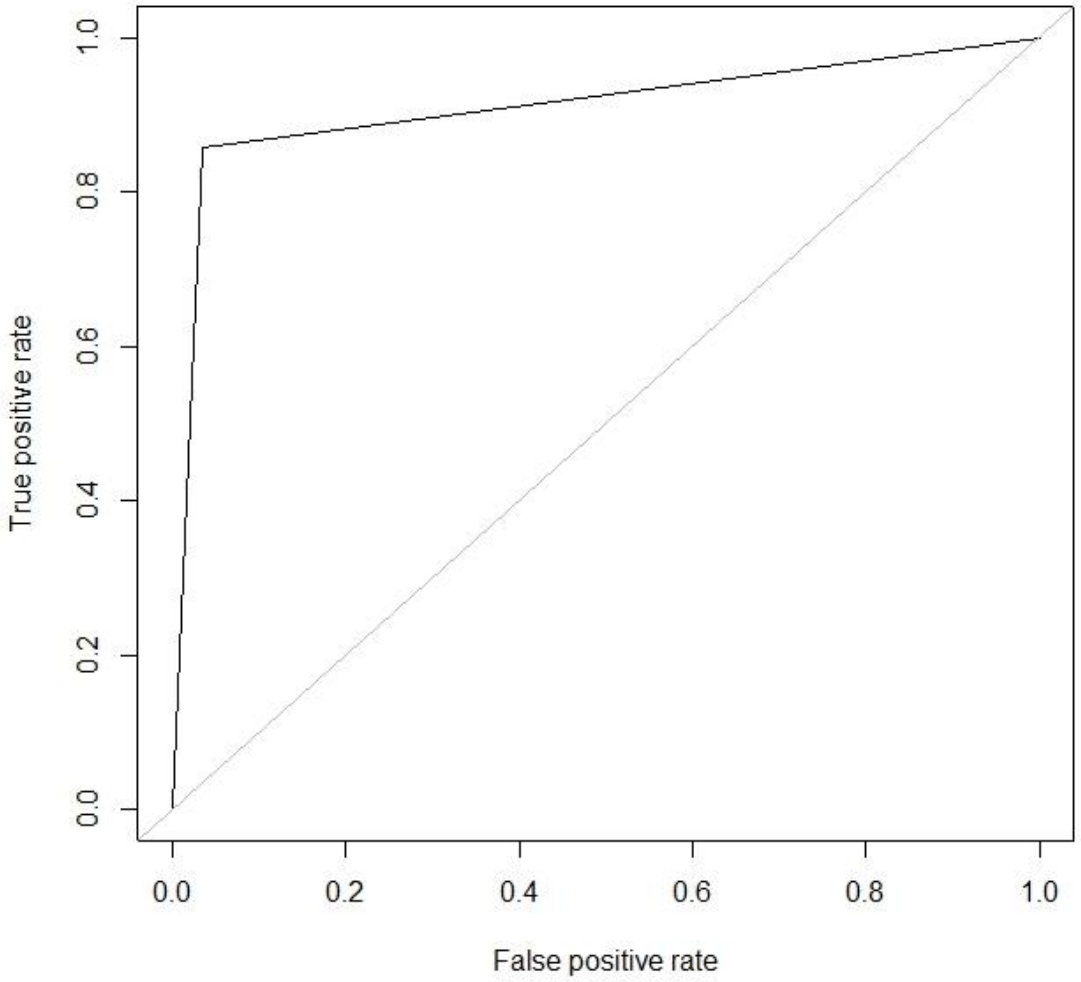


Figure 3.6 Pacific Ocean ROC: AUC = 0.91

4 CONCLUSIONS

In linear regression model, several independent variables are assumed to have direct relationship with a response variable. To overcome limitations, we developed a two-stage model method; in first stage, we use BN to evaluate relationships among selected variables in the dataset. In this stage, variables which showed direct relationship with response variables are selected. The second stage is to implement selected direct variables and response variable into a linear logistic regression model to evaluate associations. Thus, this method is simple to use as it is a recommend for association studies.

We applied this two-stage model method in a USGS dataset seeking risk/protective factors contributing a response variable: Shoreline Erosion. Six variables: smaller Tidal Range, lower Shoreline Erosion, lower Coastal Slope, higher Sea-Level Change, and higher Wave Height have higher risk-score in the CVI system. The Geomorphology Setting from rocky coast to barrier beaches; the risk score from low to high.

In the Atlantic Ocean's section, lower coastal slope (higher CVI risk score) is a direct protective factor contributing Shoreline Erosion. The more negative value in the Coastal Slope means higher risk rank-score. The Mean Wave Height showed as a directly protective factor for Shoreline Erosion. The higher Mean Wave Height has higher risk score. But, both protective factors' mechanisms for Shoreline Erosion are unknown. We recommended replication studies for these findings.

The larger Sea-Level Change and smaller Tidal Range are identified as direct risk factors contributing Shoreline Erosion. To test all of four direct risk and protective risk factors' predictive ability, we split a dataset into two datasets: training for 60% and testing for 40%. Firstly, we used parameters' estimations from the training data for the testing data to estimate

predictive values. The two protective factors (lower Coastal Slope and higher Mean Wave Height) and two risk factors (higher Sea-Level Change and smaller Tidal Range) can predict Shoreline Erosion for Atlantic Coastlines at accuracy rate 0.67 and AUC at 0.71.

In the Gulf of Mexico's section, soft Geomorphology Setting such as cobble, barrier beaches, estuary, lagoon, sand beaches, salt march, mud flats, delta, mangrove, and coral reefs, significantly directly associated with Shoreline Erosion. It consistent with correlation -0.25 between Shoreline Erosion and Geomorphology Setting. The identified direct risk factor's predictive ability for the Gulf of Mexico is at accuracy rate 0.59 and AUC at 0.63. In the Pacific's section, smaller Tidal Range as a direct protective factor is identified and associated with Shoreline Erosion. The predictive ability for the Pacific Ocean at accuracy rate at 0.98 and AUC at 0.91.

These identified direct risk/protect factors can be used for the Shoreline Erosion control; developing and evaluating a BN to calculate probabilities of long-term shoreline change. Thus, it also helps to design specific Shoreline Erosion control projects.

Limitations of this study, first is the limited resource dataset; five predictor variables: Geomorphology Setting, Sea-Level Change, Coastal Slope, Mean Tidal Range, and Mean Wave Height are not fully resourced to the discovery Shoreline Erosion. In the BN analysis, it takes two kind of variables with distributions of binomial and normal. To avoid means of transformed variables may reverse differences of means of original variables. We split either continuous or categorical variables into binary variables. The cost of dividing is: 1) the information is lost, that means the study power is reduced. 2) It may increase risk of positive results being a false positive. 3) Underestimate the extent of variation in outcome between groups. Therefore, this

could be a reason, we identified a best relationship among these variables is difference with relationship, which presented in the United States Geological Survey.

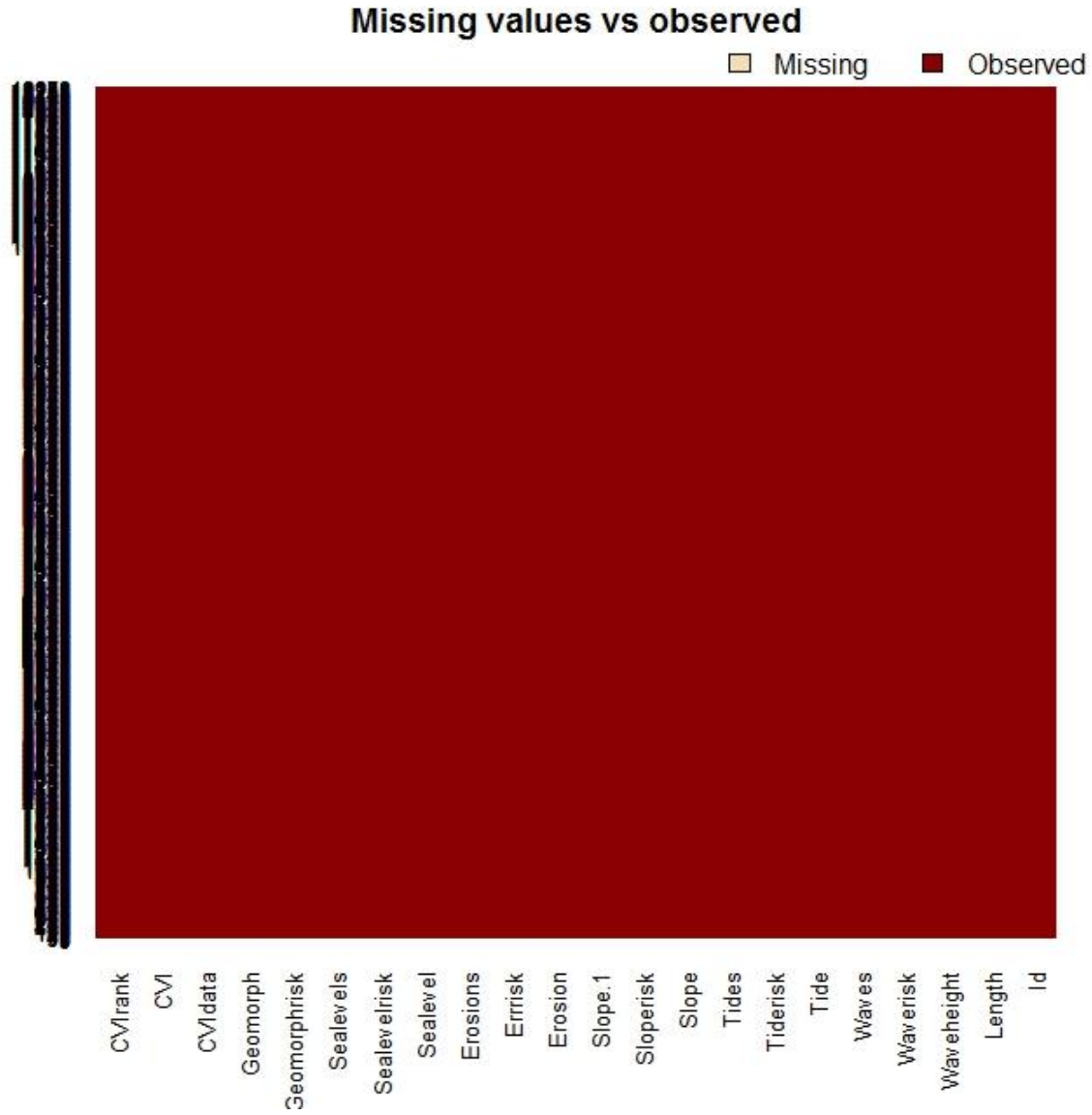
In summary, we developed a two-stage model to identify direct risk/protective factors of Mean Tidal Range, Mean Wave Height, Sea-Level Change and Coastal Slope contributing to Shoreline Erosion in the Atlantic project; Geomorphology setting and Mean Tidal Range are identified as direct risk factors for the Gulf of Mexico and Pacific projects, respectively. All of risk/protective factors are used to test predictive ability with accuracy rate ≥ 0.59 and AUC ≥ 0.63 in three US shorelines studies.

REFERENCES

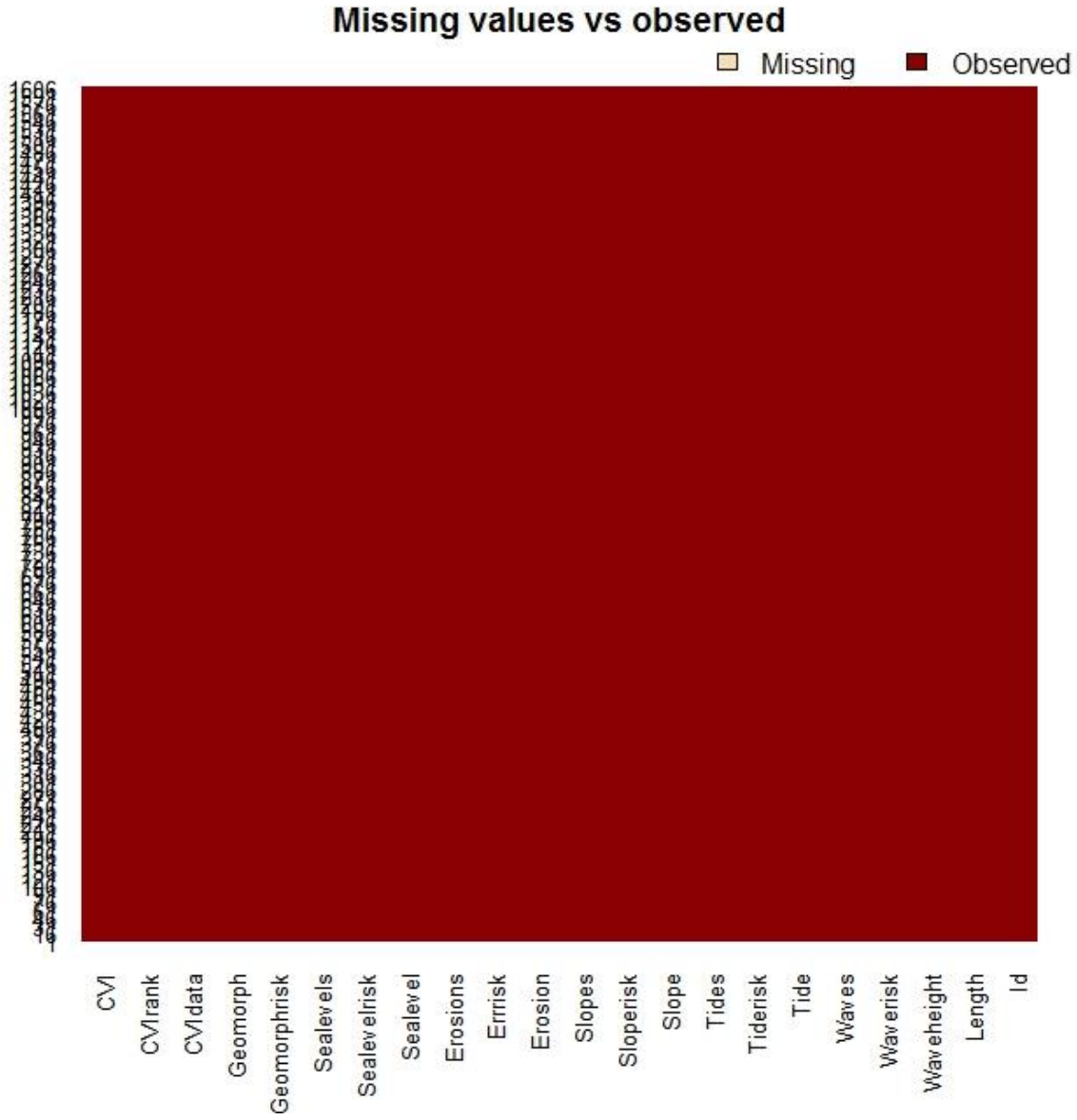
1. Thieler, E.R. and E.S. Hammar-Klose. *Vulnerability to Sea-Level Rise: Preliminary Results for the U.S. Atlantic Coast: U.S. Geological Survey Open-File Report 99-593*. 1999; Available from: http://pubs.usgs.gov/of/1999/of99-593/
2. Jensen, F.V. and T.D. Nielsen, *Bayesian networks and decision graphs* 2007, New York, NY: SpringerVerlag
3. Gutierrez, B.T., N.G. Plant, and R.E. Thieler. *A Bayesian network to predict vulnerability to sea-level rise: data report. U.S. Geological Survey Data Series 601*. 2011; Available from: https://pubs.usgs.gov/ds/601/pdf/ds601.pdf
4. Gutierrez, B.T., N.G. Plant, and E.R. Thieler, *A Bayesian network to predict coastal vulnerability to sea level rise.* Journal of Geophysical Research-Earth Surface, 2011. **116**.
5. Hammar-Klose, E.S. and E.R. Thieler. *Coastal Vulnerability to Sea-Level Rise: A Preliminary Database for the U.S. Atlantic, Pacific and Gulf of Mexico Coasts. U.S. Geological Survey Digital Data Series - 68*. 2001; Available from: http://pubs.usgs.gov/dds/dds68/htmldocs/project.htm
6. Hubertz, J.M., E.F. Thimpson, and H.V. Wang. *Wave Information Studies of US Coastlines: Annotated Bibliography on Coastal and Ocean Data Assimilation: WIS Report 36*. 1996.
7. Korb, K.B., et al., *Varieties of causal intervention.* Pricai 2004: Trends in Artificial Intelligence, Proceedings, 2004. **3157**: p. 322-331.
8. Lewis, F.I., F. Brulisauer, and G.J. Gunn, *Structure discovery in Bayesian networks: An analytical tool for analysing complex animal health data.* Preventive Veterinary Medicine, 2011. **100**(2): p. 109-115.
9. Friedman, N. and D. Koller, *Being Bayesian about network structure. A Bayesian approach to structure discovery in Bayesian networks.* Machine Learning, 2003. **50**(1-2): p. 95-125.
10. Heckerman, D., D. Geiger, and D.M. Chickering, *Learning Bayesian Networks - the Combination of Knowledge and Statistical-Data.* Machine Learning, 1995. **20**(3): p. 197-243.
11. Lewis, F.I. and B.J.J. McCormick, *Revealing the Complexity of Health Determinants in Resource-poor Settings.* American Journal of Epidemiology, 2012. **176**(11): p. 1051-1059.
12. Koivisto, M. and K. Sood, *Exact Bayesian structure discovery in Bayesian networks.* Journal of Machine Learning Research, 2004. **5**: p. 549-573.
13. Pittavino, M., F. Lewis, and R. Furrer. *R-Bayesian-networks: Additive Bayesian Network Modelling in R*. 2016; Available from: http://www.r-bayesian-networks.org/

APPENDICES

Appendix A: Atlantic Ocean Missing Data Plot

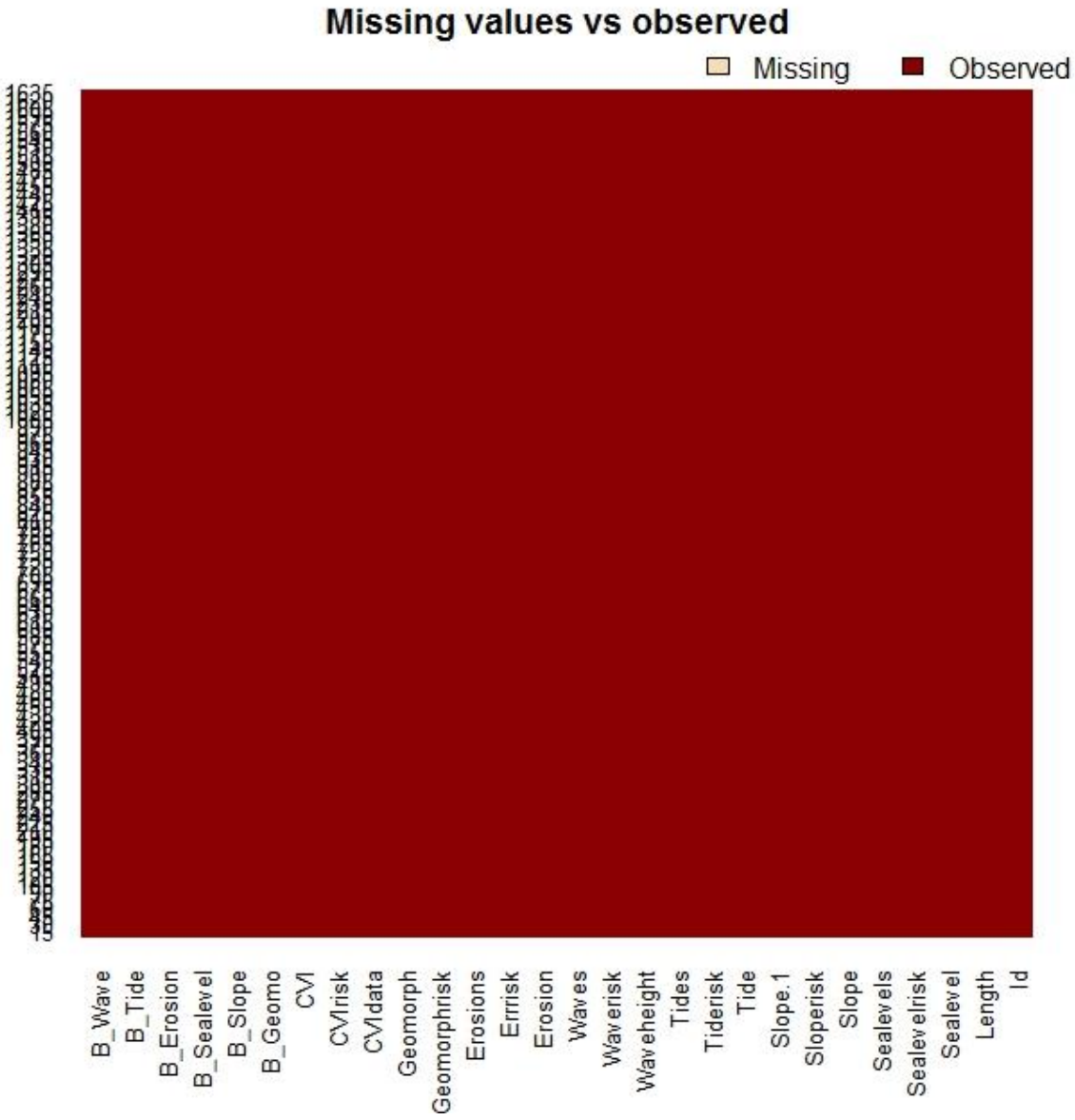


Appendix B: Gulf of Mexico Missing Data Plot



This Gulf of Mexico Data Plot has one missing value (See 2.1.3 Missing Data)

Appendix C: Pacific Ocean Missing Data Plot



Appendix D: R Code

Atlantic Ocean Shoreline Project

```
#ATLANTIC OCEAN COAST

#install.package("bnlearn")
library(bnlearn)
# install.packages("abn")
library(abn)
# install.packages("Rgraphviz")
library(Rgraphviz)
# ROC
## install.packages("pROC")
library(pROC)

setwd("H:/")
# read in data
dat <- read.csv(file="EASTCVI.csv", header=T, sep=",")
dat1 <- dat[c("Waveheight", "Tide",
"Slope","Erosion","Sealevel","Geomorphrisk")]

# correlations

# Basic Scatterplot Matrix

pairs(~Waveheight+Tide+Slope+Erosion+Sealevel+Geomorphrisk,data=
dat,main="Simple Scatterplot Matrix")
cor(dat1, use="complete.obs", method="pearson")

# check missing value
library(Amelia)
missmap(dat, main = "Missing values vs observed")
str(dat)

# basic study informaiton
quantile(dat$Length)

quantile(dat$Waveheight)
table(dat$Waverisk)/nrow(dat)

quantile(dat$Tide)
table(dat$Tiderisk)/nrow(dat)

quantile(dat$Slope)
table(dat$Sloperisk)/nrow(dat)
```

```
quantile(dat$Erosion)
table(dat$Errrisk)/nrow(dat)

quantile(dat$Sealevel)
table(dat$Sealevelrisk)/nrow(dat)

quantile(dat$Geomorphrisk)
table(dat$Geomorphrisk)/nrow(dat)

quantile(dat$CVIdata)
table(dat$CVI)/nrow(dat)

# check distributions for continuous variables
# variable distribution
# distribution of variables
par(mfrow=c(2,3))
qqnorm(dat$Geomorphrisk)
qqline(dat$Geomorphrisk)
mtext(side=3,text="Geomorph")

qqnorm(dat$Slope)
qqline(dat$Slope)
mtext(side=3,text="Slope")

qqnorm(dat$Sealevel)
qqline(dat$Sealevel)
mtext(side=3,text="Sea Level")

qqnorm(dat$Erosion)
qqline(dat$Erosion)
mtext(side=3, text = "Erosion")

qqnorm(dat$Tide)
qqline(dat$Tide)
mtext(side=3, text="Tide")

qqnorm(dat$Waveheight)
qqline(dat$Waveheight)
mtext(side=3, text = "Wave")

# variables are not normal; transfer into binary variables
# 1 = very low, low moderate, 2 = High and Higher
# make as factor

dat$B_Geomo    <- as.factor(ifelse(dat$Geomorphrisk  < 4, 1,2))
dat$B_Slope    <- as.factor(ifelse(dat$Sloperisk     < 4, 1,2))
dat$B_Sealevel <- as.factor(ifelse(dat$Sealevelrisk  < 4, 1,2))
```

```
dat$B_Erosion  <- as.factor(ifelse(dat$Errrisk       < 4, 1,2))
dat$B_Tide     <- as.factor(ifelse(dat$Tiderisk      < 4, 1,2))
dat$B_Wave     <- as.factor(ifelse(dat$Waverisk      < 4, 1,2))


#Bayesian network
dat2 <- subset(dat, select = c(B_Geomo, B_Slope, B_Sealevel ,
B_Erosion, B_Tide, B_Wave))

mydists <-
list(B_Geomo="binomial",B_Slope="binomial",B_Sealevel="binomial"
,B_Erosion="binomial",B_Tide="binomial",B_Wave="binomial")
mydag<-matrix(rep(0,36), byrow=TRUE, ncol=6)
colnames(mydag)<-rownames(mydag)<-names(dat2)

## now fit the model to calculate its goodness of fit
dat6res.c <-fitabn(dag.m = mydag, data.df = dat2,
data.dists=mydists)

## log marginal likelihood goodness of fit
print(dat6res.c)

#### Examine the parameter estimates in additive Bayesian
network

## now fit the model to calculate its goodness of fit

myres.c<-fitabn(dag.m=mydag, data.df=dat2,
data.dists=mydists,compute.fixed=TRUE)

print(names(myres.c$marginals))

#### Find the best fitting graphical structure for an additive
Bayesian network using an exact search

#use simple ban-list with no constraints

ban <- matrix(rep(0,36),byrow=TRUE,ncol=6)

colnames(ban) <-rownames(ban) <-names(dat2)

retain <- matrix(rep(0,36),byrow=TRUE,ncol=6)

colnames(retain) <-rownames(retain) <-names(dat2)
```

```
max.par <-
list("B_Geomo"=6,"B_Slope"=6,"B_Sealevel"=6,"B_Erosion"=6,"B_Tid
e"=6,"B_Wave"=6)

## now build cache
mycache <- buildscorecache(data.df=dat2,data.dists=mydists,
dag.banned=ban,dag.retained=retain,max.parents=max.par)

#now find the globally best DAG
mp.dag<-mostprobable(score.cache=mycache)

#max likelihood value
fitabn(dag.m=mp.dag,data.df=dat2,data.dists=mydists)$mlik

## plot the best model - requires Rgraphviz

### close old plot and open a new plot

plot.new()

myres<-
fitabn(dag.m=mp.dag,data.df=dat2,data.dists=mydists,create.graph
=TRUE)
plot(myres$graph)

# Final model
# B_Geomo is not directly associated wiht B_Erosion
# direct association with B_Erosion is  B_Slope
#                                       B_Sealevel
#                                       B_Tide
#                                       B_Wave

final <- glm(B_Erosion ~ B_Slope + B_Sealevel + B_Tide + B_Wave,
data = dat2, family=binomial(link="logit"))
summary(final)

## odd ratio
ORS_final <- exp(cbind(coef(final), confint(final)))
ORS_final


# ROC curve
# selected direct affected variables ( P < 0.05) in
summary(final)
# into below the model
# B_slope B_Sealevel B_Tide and B_Wave are show p < 0.05
```

```
dat3 <- subset(dat2, select = c(B_Geomo, B_Slope, B_Sealevel ,
B_Erosion, B_Tide, B_Wave))

# we randomly divided data = dat2 into two data set
# one is   train data set, random select 70%
# othe is  test  data set, 1 - 70%

sam_size <- floor(0.70 * nrow(dat3)) # 70% for test
set.seed(123456)
train_id <- sample(seq_len(nrow(dat3)), size= sam_size)
train <- dat3[train_id,  ]
test  <- dat3[-train_id, ]

# from training data set to estimate parameters

model <- glm(B_Erosion ~ B_Slope + B_Sealevel + B_Tide + B_Wave,
data = train, family=binomial(link="logit"))
summary(model)

# interpreting the results from training data set
# test null model with residual model

anova(model, test="Chisq")

#install.packages("pscl")
# no exact equivalent to r2 in linear regression
# in here, we used McFadden R2 to evaluate model fitting

library(pscl)
pR2(model)

# assessing the predictive ability of the model in test data set
# and accuracy

fitted.result <- predict(model, newdata = test, type='response')
fitted.result <- ifelse(fitted.result > 0.5, 2, 1)    # predict
0.5 as cut of value

# > 0.5 =2 , other is 1 comparing with B_Erosioin
misClassficError <- mean(fitted.result != test$B_Erosion)
print(paste("Accuracy", 1 - misClassficError))

# draw a ROC curve
#install.packages("ROCR")
library(ROCR)

### close an old plot and open a new plot
```

```
plot.new()

# predict value, it used the parameter from model in train data
set

P <- predict(model, newdata=test, type="response")

Pr <- prediction(P, test$B_Erosion)

Prf <- performance(Pr, measure = "tpr", x.measure = "fpr")
plot(Prf)
abline(c(0,0), c(1,1), col="gray")

auc <- performance(Pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

Gulf of Mexico Shoreline Project

```
#GULF OF MEXICO OCEAN COAST

#install.package("bnlearn")
library(bnlearn)
# install.packages("abn")
library(abn)
# install.packages("Rgraphviz")
library(Rgraphviz)
# ROC
## install.packages("pROC")
library(pROC)

setwd("H:/")
# read in data
dat <- read.csv(file="GULFCVI.csv", header=T, sep=",")
dat <-  na.omit(dat)

dat1 <- dat[c("Waveheight", "Tide",
"Slope","Erosion","Sealevel","Geomorphrisk")]

# correlations

# Basic Scatterplot Matrix

pairs(~Waveheight+Tide+Slope+Erosion+Sealevel+Geomorphrisk,data=
dat1,main="Simple Scatterplot Matrix")
cor(dat1, use="complete.obs", method="pearson")

# check missing value
library(Amelia)
missmap(dat, main = "Missing values vs observed")
str(dat)

# basic study information

quantile(dat$Length)

quantile(dat$Waveheight)
table(dat$Waverisk)/nrow(dat)

quantile(dat$Tide)
table(dat$Tiderisk)/nrow(dat)

quantile(dat$Slope)
table(dat$Sloperisk)/nrow(dat)
```

```
quantile(dat$Erosion)
table(dat$Errrisk)/nrow(dat)

quantile(dat$Sealevel)
table(dat$Sealevelrisk)/nrow(dat)

quantile(dat$Geomorphrisk)
table(dat$Geomorphrisk)/nrow(dat)

quantile(dat$CVIdata)
table(dat$CVI)/nrow(dat)

# check distributions for continuous variables
# variable distribution
# distribution of variables
par(mfrow=c(2,3))
qqnorm(dat$Geomorphrisk)
qqline(dat$Geomorphrisk)
mtext(side=3,text="Geomorph")

qqnorm(dat$Slope)
qqline(dat$Slope)
mtext(side=3,text="Slope")

qqnorm(dat$Sealevel)
qqline(dat$Sealevel)
mtext(side=3,text="Sea Level")

qqnorm(dat$Erosion)
qqline(dat$Erosion)
mtext(side=3, text = "Erosion")

qqnorm(dat$Tide)
qqline(dat$Tide)
mtext(side=3, text="Tide")

qqnorm(dat$Waveheight)
qqline(dat$Waveheight)
mtext(side=3, text = "Wave")

# variables are not normality, transfer into binary variables
# 1 = very low, low moderate, 2 = High and very Higher
# make as factor
dat$B_Geomo    <- as.factor(ifelse(dat$Geomorphrisk  < 4, 1,2))
dat$B_Slope    <- as.factor(ifelse(dat$Sloperisk     < 4, 1,2))
dat$B_Sealevel <- as.factor(ifelse(dat$Sealevelrisk  < 4, 1,2))
```

```
dat$B_Erosion  <- as.factor(ifelse(dat$Errrisk      < 4, 1,2))
dat$B_Tide     <- as.factor(ifelse(dat$Tiderisk      < 4, 1,2))
dat$B_Wave     <- as.factor(ifelse(dat$Waverisk      < 4, 1,2))


#Bayesian network
dat2 <- subset(dat, select = c(B_Geomo, B_Slope, B_Sealevel ,
B_Erosion, B_Tide, B_Wave))

mydists <-
list(B_Geomo="binomial",B_Slope="binomial",B_Sealevel="binomial"
,B_Erosion="binomial",B_Tide="binomial",B_Wave="binomial")
mydag<-matrix(rep(0,36), byrow=TRUE, ncol=6)
colnames(mydag)<-rownames(mydag)<-names(dat2)

## now fit the model to calculate its goodness of fit
dat6res.c <-fitabn(dag.m = mydag, data.df = dat2,
data.dists=mydists)

## log marginal likelihood goodness of fit
print(dat6res.c)

#### Examine the parameter estimates in additive Bayesian
network

## now fit the model to calculate its goodness of fit

myres.c<-fitabn(dag.m=mydag, data.df=dat2,
data.dists=mydists,compute.fixed=TRUE)

print(names(myres.c$marginals))

#### Find the best fitting graphical structure for an additive
Bayesian network using an exact search

#use simple banlist with no constraints

ban <- matrix(rep(0,36),byrow=TRUE,ncol=6)

colnames(ban) <-rownames(ban) <-names(dat2)

retain <- matrix(rep(0,36),byrow=TRUE,ncol=6)

colnames(retain) <-rownames(retain) <-names(dat2)
```

```
max.par <-
list("B_Geomo"=6,"B_Slope"=6,"B_Sealevel"=6,"B_Erosion"=6,"B_Tid
e"=6,"B_Wave"=6)

## now build cache
mycache <- buildscorecache(data.df=dat2,data.dists=mydists,
dag.banned=ban,dag.retained=retain,max.parents=max.par)

#now find the globally best DAG
mp.dag<-mostprobable(score.cache=mycache)

#max likelihood value
fitabn(dag.m=mp.dag,data.df=dat2,data.dists=mydists)$mlik

## plot the best model - requires Rgraphviz

### since an original plot has six small plots and close an old
plot
### and open a new plot

plot.new()

myres<-
fitabn(dag.m=mp.dag,data.df=dat2,data.dists=mydists,create.graph
=TRUE)
plot(myres$graph)

# Final model
# B_Wave, and B_Slope are not directly associated wiht B_Erosion
# direct association with B_Erosion is  B_Selevel
#                                       B_Geomo
#                                       B_Tide

final <- glm(B_Erosion ~ B_Geomo + B_Sealevel + B_Tide, data =
dat2, family=binomial(link="logit"))
summary(final)

## odd ratio
ORS_final <- exp(cbind(coef(final), confint(final)))
ORS_final

# ROC curve
# selected direct affected variables ( P < 0.05) in
summary(final)
# into below the model
# B_Geomo < 0.05
```

```
dat3 <- subset(dat2, select = c(B_Geomo,B_Erosion))

# we randomly divided data = dat2 into two data set
# one is   train data set, random select 70%
# other is  test  data set, 1 - 70%

sam_size <- floor(0.70 * nrow(dat3)) # 70% for test
set.seed(123456)
train_id <- sample(seq_len(nrow(dat3)), size= sam_size)
train <- dat3[train_id,  ]
test  <- dat3[-train_id, ]

# from training data set to estimat parameters

model <- glm(B_Erosion ~ B_Geomo, data = train,
family=binomial(link="logit"))
summary(model)

# interpreting the results from training data set
# test null model with residual model

anova(model, test="Chisq")

#install.packages("pscl")
# no exact equivalent to r2 in linear regression
# in here, we used McFadden R2 to ecaluate model fitting

library(pscl)
pR2(model)

# assessing the predictive ability of the model in test data set
# and accuracy

fitted.result <- predict(model, newdata = test, type='response')
fitted.result <- ifelse(fitted.result > 0.5, 2, 1)    # predict
0.5 as cut of value
# > 0.5 =2 , other is 1 comparing with B_Erosioin
misClassficError <- mean(fitted.result != test$B_Erosion)
print(paste("Accuracy", 1 - misClassficError))

#draw ROC curve
#install.packages("ROCR")
library(ROCR)

# predict value, it used the parameter from model in train data
set
```

```
P <- predict(model, newdata=test, type="response")

Pr <- prediction(P, test$B_Erosion)

Prf <- performance(Pr, measure = "tpr", x.measure = "fpr")
plot(Prf)
abline(c(0,0), c(1,1), col="gray")

auc <- performance(Pr, measure = "auc")
auc <- auc@y.values[[1]]
auc
```

Pacific Ocean Shoreline Project

```
#PACIFIC OCEAN COAST

#install.package("bnlearn")
library(bnlearn)
# install.packages("abn")
library(abn)
# install.packages("Rgraphviz")
library(Rgraphviz)
# ROC
## install.packages("pROC")
library(pROC)

setwd("H:/")
# read in data
dat <- read.csv(file="PACCVI.csv", header=T, sep=",")
dat1 <- dat[c("Waveheight", "Tide",
"Slope","Erosion","Sealevel","Geomorphrisk")]

# correlations

# Basic Scatterplot Matrix

pairs(~Waveheight+Tide+Slope+Erosion+Sealevel+Geomorphrisk,data=
dat,main="Simple Scatterplot Matrix")
cor(dat1, use="complete.obs", method="pearson")

# check missing value
library(Amelia)
missmap(dat, main = "Missing values vs observed")
str(dat)

# basic study information

quantile(dat$Geomorphrisk)
table(dat$Geomorphrisk)/nrow(dat)

quantile(dat$Slope)
table(dat$Sloperisk)/nrow(dat)

quantile(dat$Sealevel)
table(dat$Sealevelrisk)/nrow(dat)

quantile(dat$Erosion)
table(dat$Errrisk)/nrow(dat)
```

```
quantile(dat$Tide)
table(dat$Tiderisk)/nrow(dat)

quantile(dat$Waveheight)
table(dat$Waverisk)/nrow(dat)

# check distributions for continues varialbes
# variable distribution
# distribution of variables
par(mfrow=c(2,3))
qqnorm(dat$Geomorphrisk)
qqline(dat$Geomorphrisk)
mtext(side=3,text="Geomorph")

qqnorm(dat$Slope)
qqline(dat$Slope)
mtext(side=3,text="Slope")

qqnorm(dat$Sealevel)
qqline(dat$Sealevel)
mtext(side=3,text="Sea Level")

qqnorm(dat$Erosion)
qqline(dat$Erosion)
mtext(side=3, text = "Erosion")

qqnorm(dat$Tide)
qqline(dat$Tide)
mtext(side=3, text="Tide")

qqnorm(dat$Waveheight)
qqline(dat$Waveheight)
mtext(side=3, text = "Wave")

# variables are not normality, transfer into binary variables
# 1 = very low, low moderate, 2 = High and very Higher
# make as factor

dat$B_Geomo    <- as.factor(ifelse(dat$Geomorphrisk  < 4, 1,2))
dat$B_Slope    <- as.factor(ifelse(dat$Sloperisk     < 4, 1,2))
dat$B_Sealevel <- as.factor(ifelse(dat$Sealevelrisk  < 4, 1,2))
dat$B_Erosion  <- as.factor(ifelse(dat$Errrisk      < 4, 1,2))
dat$B_Tide     <- as.factor(ifelse(dat$Tiderisk      < 4, 1,2))
dat$B_Wave     <- as.factor(ifelse(dat$Waverisk      < 4, 1,2))


#Bayesian network
```

```
dat2 <- subset(dat, select = c(B_Geomo, B_Slope, B_Sealevel ,
B_Erosion, B_Tide, B_Wave))

mydists <-
list(B_Geomo="binomial",B_Slope="binomial",B_Sealevel="binomial"
,B_Erosion="binomial",B_Tide="binomial",B_Wave="binomial")
mydag<-matrix(rep(0,36), byrow=TRUE, ncol=6)
colnames(mydag)<-rownames(mydag)<-names(dat2)

## now fit the model to calculate its goodness of fit
dat6res.c <-fitabn(dag.m = mydag, data.df = dat2,
data.dists=mydists)

## log marginal likelihood goodness of fit
print(dat6res.c)

#### Examine the parameter estimates in additive Bayesian
network

## now fit the model to calculate its goodness of fit

myres.c<-fitabn(dag.m=mydag, data.df=dat2,
data.dists=mydists,compute.fixed=TRUE)

print(names(myres.c$marginals))

#### Find the best fitting graphical structure for an additive
Bayesian network using an exact search

#use simple banlist with no constraints

ban <- matrix(rep(0,36),byrow=TRUE,ncol=6)

colnames(ban) <-rownames(ban) <-names(dat2)

retain <- matrix(rep(0,36),byrow=TRUE,ncol=6)

colnames(retain) <-rownames(retain) <-names(dat2)

max.par <-
list("B_Geomo"=6,"B_Slope"=6,"B_Sealevel"=6,"B_Erosion"=6,"B_Tid
e"=6,"B_Wave"=6)

## now build cache
mycache <- buildscorecache(data.df=dat2,data.dists=mydists,
dag.banned=ban,dag.retained=retain,max.parents=max.par)
```

```
#now find the globally best DAG
mp.dag<-mostprobable(score.cache=mycache)

#max likelihood value
fitabn(dag.m=mp.dag,data.df=dat2,data.dists=mydists)$mlik

## plot the best model - requires Rgraphviz

### close an old plot then make a new plot

plot.new()

myres<-
fitabn(dag.m=mp.dag,data.df=dat2,data.dists=mydists,create.graph
=TRUE)
plot(myres$graph)

# Final model
# B_Sealevel is not directly associated wiht B_Erosion
# direct associated with B_Erosion is  B_Slope
#                                      B_Gemo
#                                      B_Tide
#                                      B_Wave

final <- glm(B_Erosion ~ B_Slope + B_Geomo + B_Tide + B_Wave,
data = dat2, family=binomial(link="logit"))
summary(final)

## odd ratio
ORS_final <- exp(cbind(coef(final), confint(final)))
ORS_final

# ROC curve
# selected direct affected variables ( P < 0.05) in
summary(final)
# into below the model
# B_tidep < 0.05

dat3 <- subset(dat2, select = c(B_Erosion, B_Tide))

# we randomly divided data = dat2 into two data set
# one is   train data set, random select 70%
# othe is  test  data set, 1 - 70%

sam_size <- floor(0.70 * nrow(dat3)) # 70% for test
set.seed(123456)
train_id <- sample(seq_len(nrow(dat3)), size= sam_size)
```

```
train <- dat3[train_id,  ]
test  <- dat3[-train_id, ]

# from training data set to estimate parameters

model <- glm(B_Erosion ~ B_Tide , data = train,
family=binomial(link="logit"))
summary(model)

# interpreting the results from training data set
# test null model with residual model

anova(model, test="Chisq")

#install.packages("pscl")
# no exact equivalent to r2 in linear regression
# in here, we used McFadden R2 to evaluate model fitting

library(pscl)
pR2(model)

# assessing the predictive ability of the model in test data set
# and accuracy

fitted.result <- predict(model, newdata = test, type='response')
fitted.result <- ifelse(fitted.result > 0.5, 2, 1)    # predict
0.5 as cut of value
# > 0.5 =2 , other is 1 comparing with B_Erosioin
misClassficError <- mean(fitted.result != test$B_Erosion)
print(paste("Accuracy", 1 - misClassficError))

# drwo ROC curve
#install.packages("ROCR")
library(ROCR)

# predict value, it used the parameter from model in train data
set

P <- predict(model, newdata=test, type="response")

Pr <- prediction(P, test$B_Erosion)

Prf <- performance(Pr, measure = "tpr", x.measure = "fpr")
plot(Prf)
abline(c(0,0), c(1,1), col="gray")

auc <- performance(Pr, measure = "auc")
```

```
auc <- auc@y.values[[1]]
auc
```