

A Bayesian Approach to Approximating Onshore Ocean Wave Models with First Person Blog Reports

Bilson Jake Libres Campana
Department of Computer Science and Engineering
University of California Riverside
900 University Ave
Riverside, CA, USA, 92521
bcampana@cs.ucr.edu

ABSTRACT

In recent years, wave forecasts have improved significantly both in empirical and spectral approaches. These methods utilize information such as mid-ocean wave heights, wind sea and swell, mean wave direction and directional wave spectra in order to create global models of our ocean's surface. To accurately predict onshore waves, shore local features must be taken into account in the prediction model. Documentation of these features is close to an impossible task not only due to the sheer amount of the world's coasts, about 217,000 miles (roughly the distance from the earth to the moon), but also to the dynamic nature of these features. As with the process of facing any seemingly impossible calculation, I propose to approximate onshore wave models by building a Bayesian network which considers first person reports of wave heights at beaches.

1. INTRODUCTION

Wave forecasts have improved significantly both in empirical [9] and spectral [1] approaches. These models also take into account bathymetric, which specify ocean depths and land masses, as well as obstruction charts which account for uncharted islands (which may be dynamically produced through volcanic activity) and mobile or growing ice masses [10]; however, even accounting for such local variations cannot accurately predict onshore wave models to an acceptable accuracy. To accurately predict onshore waves, shore local features must be taken into account in the prediction model. These features may include manmade structures, jetties, septic pipes, piers; natural formations such as sand bars, reefs, bays; or even reoccurring events such as canal openings or docked crafts. The quality and structure of these reports vary from person to person and it would be assumed that a higher quality report would result in a higher quality approximation. Towards this I have utilized the surf reports of [11] as the first person report input. The Bayesian model will consider the common variables which many models consider [3] as well as new variables which may have influence onshore local variations (i.e. inland rain

which could cause ocean run off that modifies the onshore ocean floor).



Figure 1 - (Above) Large map showing sensor locations. (Below) Zoomed map of Costa Rica. Locations are: (1) Una Loa Resort, (2) Liberia, (3) Tobias Bolanos Airport, (4) Puerto Limon, (5) Buoy 43412, (6) Buoy 33411, and (7) Buoy 32315

The following paper will discuss the data used in our approach in section 2, a brief overview of Bayesian Networks, graph scoring, and structure learning in sections 3, 4, and 5 respectively, experimentation and analysis in section 6, a discussion of current and future work in section 7, and concludes in section 8.

2. DATA

Data for this project was gathered from three sources: Meteorological Assimilation Data Ingest System (MADIS)[6], the National Data Buoy Center (NBDC)[8], and the Una Ola resort[11]. Figure 1 displays the geographic locations of the sensors and Table 1 summarizes the variables and their discretizations. Observations begin from January 4, 2009 to December 3, 2009 for a total of 317 observations.

Variable Name (Symbol)	Values
Onshore Wave Height (H)	3,6, or 10
Water Column Height – Panama (PH)	2
Water Column Height – Manzanillo (MH)	2
Zonal Winds (OZ)	2
Meridional Winds (OM)	2
Wind Speed (OS)	2
Wind Direction (OD)	4
Air Temperature (OT)	2
Sub-surface Water Temperature (OW)	2
Relative Humidity (OR)	2
Liberia Air Temperature (LT)	2
Liberia Dew Point (LD)	2
Liberia Wind Speed (LS)	2
Puerto Limon Air Temperature (PT)	2
Puerto Limon Dew Point (PD)	2
Puerto Limon Wind Speed (PS)	2
Tobias Bolanos INTL Air Temperature (TT)	2
Tobias Bolanos Dew Point (TD)	2
Tobias Bolanos Wind Speed (TS)	2

Table 1 - List of model variables and value discretizations

2.1 Sensor and Variable Information

2.1.1 Costa Rica

Data was observed by a local surf resort. The resort posts daily surf reports at their location onto a blog.

Onshore wave height (H)

Normally, surf reports are given by a range of wave heights (i.e. "... with waves 3-5 ft..."). When surfers use these ranges, the trend is to give the most likely of the largest of wave sizes as an upper bound. Larger wave heights are given as outliers and are usually reported separately using the word "occasional" (i.e. "... with an occasional 8 ft. wave..."). The maximum of each range was taken into data for this project. This variable has been discretized into 3, 6, and the original 10 possible heights.

2.1.2 Oceanic Buoys 43412 and 33411

These data were recorded from an oceanic buoy offshore from Manzanillo, Mexico and Panama City, Panama at nominal coordinates 15N, 105W and 5N, 90W respectively.

Water column height (MH,PH)

These variables are the daily average heights of the water column in meters at the Manzanillo and Panama buoys re-

spectively. High averaging heights give a higher possibility for high onshore waves.

2.1.3 Oceanic Buoy 32315

This buoy is moored at 5N, 110W and is outfitted with more sensors allowing it to gather significantly more information than the previous buoys.

Zonal Winds (OZ)

The average wind speed of zonal winds (winds parallel to the latitude) measured in meters per second. High wind speed allow for larger wind swells (waves generated by the movement of winds). A zonal wind contributes to a West-East swell.

Meridional Winds (OM)

The average wind speed of zonal winds (winds parallel to the longitude) measured in meters per second. A meridional wind contributes to a North-South swell.

Wind Speed (OS)

The average daily wind speed measured in meters per second.

Wind Direction (OD)

The average daily wind direction measured in degrees.

Air Temperature (OT)

Average air temperature measured in degrees Celsius.

Sub-surface Water Temperature (OW)

The average sub-surface water temperature measured in degrees Celsius.

Relative Humidity (OR)

The average relative humidity measured in percentages. Humidity of the air is a measure of its saturation with water. High humidity increases the probability of condensation, rather than evaporation, which may lead to rain and thunders-torms.

2.1.4 Liberia, Puerto Limon, Tobias Bolanos

International Airport

These measurements were taken at land based weather stations in Liberia, Puerto Limon, and the Tobias Bolanos International Airport.

Temperature (LT, PT, TT)

The average daily air temperature measured in degrees Fahrenheit.

Dew Point (LD, PD, TD)

The average daily dew point temperature measured in degrees Fahrenheit. The dew point is the temperature air must be cooled for it to condense into water. It is associated with the relative humidity. High humidity means the dew point is closer to the current air temperature.

Wind Speed (LS, PS, TS)

The average daily wind speed measured in meters per second.

2.2 DISCRETIZATION

For computational and representational simplicity, each of the variables is discretized into equal sized bins of a zero mean, unit standard deviation normal distribution. Each variable in the data set is Z-normalized and then binned against this distribution.

3. BAYESIAN NETWORKS

We learn Bayesian networks to represent our model and for inference. We briefly introduce pertinent topics of Bayesian networks, a full reference can be found in [4].

3.1 Representation

A Bayesian network graph is represented by a directed acyclic graph (DAG). A directed edge between two nodes signifies that they interact directly with each other and could be dependent. Each node in the graph represents a variable in the model and each contains a probability distribution which includes variables with edges that input into it. This distribution may be in the form of a conditional probability table, in the case of discrete variables, or it can be approximated with a function such as a Gaussian, in the case of continuous variables.

3.2 Inference

With a Bayesian network we can infer the probability of events, given some evidence, without calculating the entire joint distribution of all variables. The dependency information which the graph indicates allows us to decrease the amount of calculations necessary by allowing us to ignore variables which are independent of our variables of interest given that we have some evidence set.

In our project, our variable of interest is the onshore wave height (H). We would also like to minimize the calculations and variables necessary to achieve this.

3.3 Learning

Constructing a graph can require the knowledge of an expert in the field. If our graph has a large number of variables with high dimensionality, it may be infeasible to ask someone to construct a graph based on them. It can also be difficult to extract independency information from experts. Many times our knowledge or intuition is difficult to explain explicitly to others, it simply makes sense to us from our experiences and background.

To avoid these complications we can utilize methods to learn the structure of a graph from data. We would like this graph to represent a distribution P which closely approximates the true distribution P^* . Algorithms such as greedy search and minimum weighted spanning tree are discussed in section 5.

4. GRAPH SCORING

Given a graph, we'd like to have a heuristic to tell us how well it represents the underlying distribution. Specifically, we'd like measure how much the independence assumptions in the approximated distribution differ from the assumptions in the underlying distribution. Ideally, we would like:

$$P^*(X, Y) = P^*(X)P^*(Y) = P(X)P(Y) = P(X, Y).$$

If we do not have access to the true distribution, then we have to assume that the counts in our data $M[X, Y]$ are our best approximation \hat{P} . So that:

$$M * P^*(X)P^*(Y) = M[X, Y] = M * \hat{P}(X)\hat{P}(Y).$$

Given this assumed approximation, we can now measure the deviation of a graph from the approximated underlying distribution.

4.1 Mutual Information

Mutual information is a measure of the mutual dependence of two variables.

$$I(X; Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}.$$

If $I(X; Y) = 0$, then X and Y are independent and observing one variable tells us nothing of the other variable. Because this measure is symmetric, we cannot distinguish between the graph $X \rightarrow Y$ and $Y \rightarrow X$. This score can then only favor a family of I-equivalent graphs with a maximal score for the data. This property of the score is known as *score equivalence*.

4.2 Bayesian Information Criterion

The Bayesian information criterion (BIC) score measures the likelihood of the graph with our data with a penalty for the complexity of graph G .

$$score_{BIC}(G; D) = M \sum_{i=1}^n I(X_i; Pa_{X_i}^G) - M \sum_{i=1}^n H(X_i) - \frac{\log M}{2} Dim[G].$$

Here, M is the cardinality of our data, $Dim[G]$ is the number of parameters in our graph, $H(X_i)$ is the entropy of node X_i :

$$H(X) = \sum_{x \in X} P(x) \log \frac{1}{P(x)},$$

and $Pa_{X_i}^G$ is the parents of node X_i in graph G . This score also exhibits score equivalence.

5. STRUCTURE SEARCH

Now that we have a measure to score graphs against a distribution, we can now conduct searches to find an optimal network to fit our data. This project utilizes the Bayesian network toolboxes provided by [5] and [7].

5.1 Greedy Search

A greedy algorithm attempts to approximate the global optimum result by following locally optimal decisions. With structure search of Bayesian networks, we atomize the possible operations $o(G)$ from a given network graph G to produce a new graph G' . These operations are:

1. Add an edge.
2. Remove an edge.
3. Reverse an edge.

With these operations we can define a set of graphs G_o which are the graphs that are one operation away from our current graph. Finding the maximum scoring graph of this set tells us what direction to move in the current stage of our search. We can continue this greedy search (GS) until all possible neighbors of graph provide a score lower than that of our current graph.

5.1.1 Constrained Hill Climbing (CHC) Search

Because our gathered data is being used to approximate the underlying distribution, an insufficient amount of data would give an inaccurate representation. We may then like to impose some prior probabilities on the possible graphs to account for this inaccuracy. In our model, we know that the weather and sea level heights of the waves at proximate points affect the onshore wave model. But without enough data the onshore wave heights may seem independent of these factors. Searches on the data could then turn up graphs

where the onshore wave height variable is independent from the rest of the graph. To keep this from occurring we apply a constraint which does not allow the greedy search algorithm to take a step that would produce a graph with no connectivity to the variable H.

5.1.2 CHC with Random Restarts (CHC+RR)

Because of our search dynamics, it is possible to get stuck at local maxima in the search space. To attempt to find other local maxima we perform CHC 10 times starting with random graph configurations.

5.2 Maximum Weight Spanning Tree

A maximum weight spanning tree is a tree structure which spans all vertices and whose sum of edge weights is maximum when compared to all other possible spanning trees of the graph. We use the maximum weight spanning tree (MWST) algorithm in this project with two scoring functions for the edge weights: mutual information and the BIC score. This algorithm creates a graph by initially considering a single entry node set of the root node, adding the highest scoring edge which originates from our current node set to the remaining nodes, added the destination node to our node set, and repeating until our node set contains all nodes. This algorithm is run once with each node as a root for an exhaustive root search (ERS).

6. EXPERIMENTAL RESULTS AND ANALYSIS

NS Parameters	NS3, NS6, NS10
Structure Algorithms	GS, CHC, MWST+ERS, CHC+RR
Seed Graphs	ZC, DH, FC
MWST Score Functions	MI, BIC

Table 2 – Summary of possible values for configurations in our experiments.

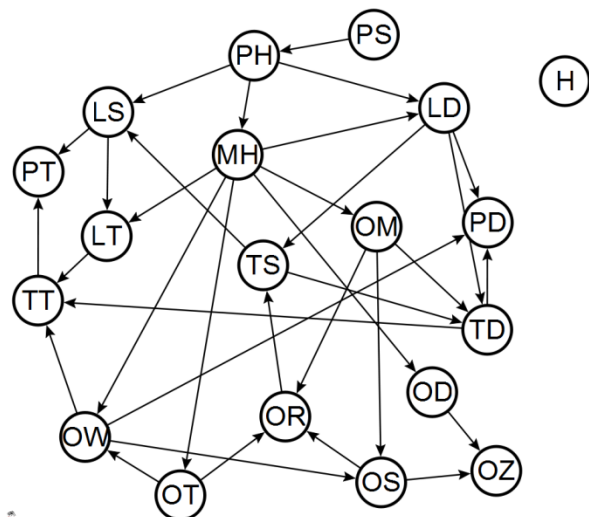


Figure 2 - Example of an unconstrained GS graph. H is consistently found independent.

As noted in Table 1, three discretizations of the data have been used with each varying on the value cardinality of the

onshore wave height variable H: NS3, NS6, and NS10 with a node size of variable H of three, six, and ten respectively. The GS algorithm is initially run unconstrained with three different connectivity seed graphs: zero connectivity (ZC), all dependent on variable H (DH), and full connectivity (FC). The GS algorithm is also run in its constrained form as the CHC algorithm. CHC is also run with all three seed graphs. The MWST algorithm is ran on the three NS parameters using two score functions of edge weight calculations: mutual information (MI) and the Bayesian Information Criterion (BIC) score. An exhaustive root search (ERS) is done to find the highest scoring MWST graph. All networks are compared using the BIC score. These settings are summarized in Table 2.

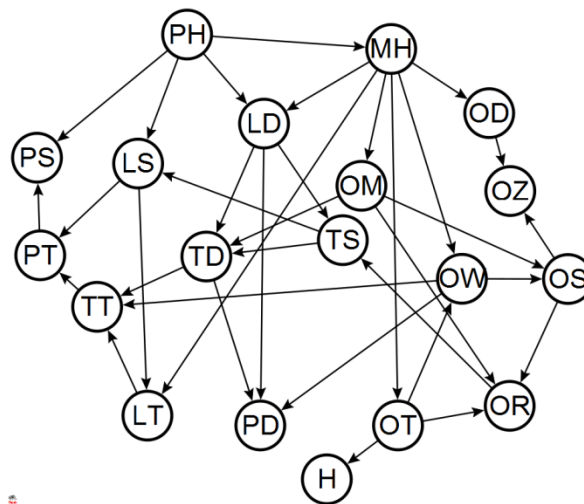


Figure 3 – Best network with NS3 parameters. Found with CHC – no connectivity.

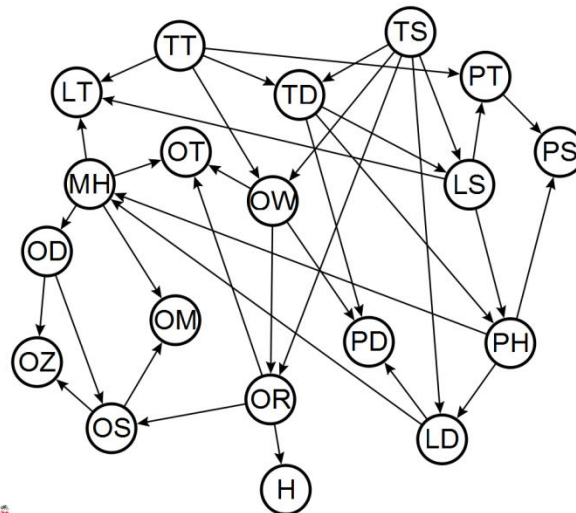


Figure 4 – Best network with NS6 parameters. Found with CHC – one to all connectivity.

Searching with the GS algorithm continuously produced DAGs which had variable H independent from the rest of the graph, such as the graph in Figure 2. This is caused by the limited amount of observations we learned our structure with. We know that the onshore wave height H is, at least,

dependent on the mid ocean measurements. Using this knowledge we run the CHC algorithm in order to find a network with a dependent H variable.

Figure 3, Figure 4, and Figure 5 each show the highest scoring network from the CHC structure search algorithm using the NS parameters NS3, NS6, and NS10 respectively. These graphs shows that the onshore wave height is consistently dependent on our mid ocean readings of relative humidity and temperature. Intuitively, we can believe there might be a dependence on the ocean temperature. Large changes in ocean air temperatures can signify seasonal changes. Larger wave heights are more commonly associated with colder weathers. This is due to the possible causes which occur during colder seasons such as lower pressure systems and storms. These causes can also imply the presence of strong winds which increases the possibility of large wind swells.

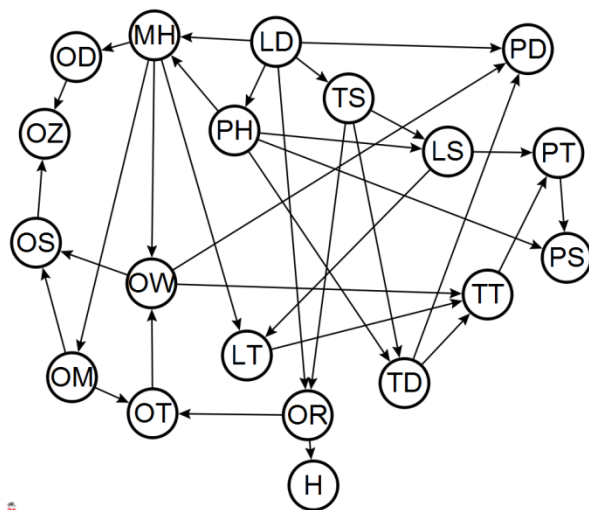


Figure 5 – Best parameters with NS10 parameters. Found with CHC – one to all connectivity.

The dependence on mid ocean humidity is less obvious. Relative humidity (RH) is a measure of how much the air is saturated with water [12]. High RH values imply that there is a large amount of water vapor in the atmosphere. When this happens it is more likely for condensation of water to happen than its evaporation. Humid air will rise up into the atmosphere and mix with less humid air. This is the natural process for the creation of thunderstorms and other weather phenomena. The increased likelihood of storms increases the possibility of wind and wind swell, which increases the probability of high onshore waves.

Just like the CHC algorithm, the MWST graphs, such as the graph in Figure 6, also assigned the variable H dependent on either OT or OR. These graphs were very simple in that they exhibited a large amount of node chains. Using these chained graphs, a single variable in evidence could dramatically lower the amount of calculations necessary during inference.

The CHC+RR algorithm did not always found a higher scoring graph than a single run CHC. The CHC+RR algorithm ran in each of the NS settings found between two to six local maxima. This shows that our search space is highly modal, making it difficult to find a global maximum using a hill climbing search.

Besides focusing on our variable of interest of H, our models show us other interesting dependencies. In each of the graphs, it can be seen that there is a dependency between the onshore temperatures of Puerto Limon, Tobias INTL, and Liberia. This is not a surprising dependency since the country is relatively small.

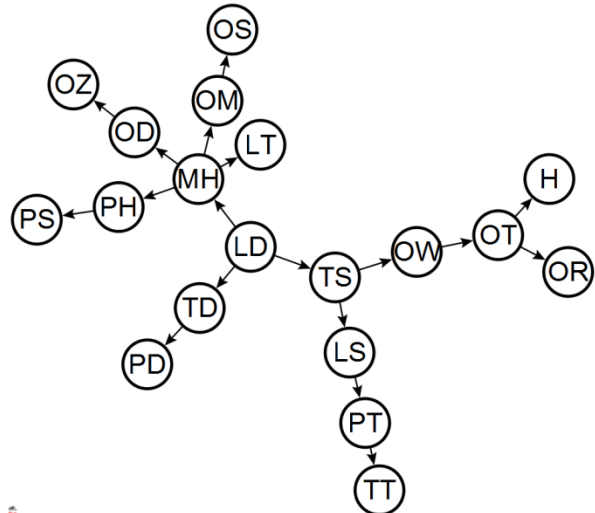


Figure 6 - MWST graph with NS3 parameters and BIC score. Best root with all parameters is with LD.

In the MWST of Figure 6, the temperature of Liberia (LT) can be independent of the other two temperatures (TT,PT) given some variables, which include wind speeds at Liberia (LS) and Tobias INTL (TS) as well as the dew point at Liberia (LD). This is because Costa Rica is part of the Pacific Rim of Fire [2], which riddles the country with large mountain ranges and volcanoes. Separating the Liberia from Puerto Limon and Tobias INTL is the Tilaran mountain range. High dew points may be an indicator of storms. Storms and large air speeds raise the chances of having a large storm system that may cover multiple regions. Where on a normal day, without their presence, the mountain range can cause a separation of the region's climates. The graphs also depicts the obvious dependency of the zonal wind speeds (OZ) to the direction of the wind (OD) and its speed (OS).

7. DISCUSSION AND FUTUREWORK

There were difficulties in collecting data because the data collection standards of the region were not up to those of the USA. Many times a sensor would go down for several months. There were also intermittent blackouts of data with other sensors. An extension of this work would be to include unobserved variables or missing data. Now that we have a better understanding of the network, we can create a seed graph for algorithms such as structural EM which deal with missing data and variables. Another approach would be to select a much more trafficked location with a more dependable array of sensors and data. The National Oceanic and Atmospheric Administration helps maintain a thorough network of weather sensors which could be used for popular shore lines and ports in the USA. I have contacted popular websites which regularly publish surf reports, but have not been able to acquire data from them. Acquiring such data would give us a valuable source of first person reports from a large net-

work of reporters from several years ago. Increasing the samples would help increase our chances of finding a graph that approximates the underlying distribution better.

8. CONCLUSION

Calculating onshore ocean wave heights with current approaches is an infeasible task due to the variability of shores and the thousands of miles of shoreline on our planet. This Bayesian network approach allows us to approximate this by considering meteorological data along with site specific first person accounts. By considering these reports we imply features of the shore such as sand bars, river mouths, and piers. This method also allows us to use freely available data and lets us avoid time consuming surveys of the local environment and topology to construct our dependency graphs. These graphs also allow a non expert to understand some of the dependencies of meteorological measurements in the area. Within our data, the models we have found have shown us that we only need to consider a single observed variable in order to get a distribution over the onshore wave heights. This approach not only gave us a distribution of our variable of interest but also offered a dramatic reduction in our dimensionality.

9. REFERENCES

- [1] Bretschneider, C. L., 1958: Revisions in wave forecasting: Deep and shallow water. *Proc. 6th Int. Conf. Coastal Eng.*, ASCE, 30-67.
- [2] CentralAmerica.Com: Costa Rica Information.
<http://centralamerica.com/cr/info/>
- [3] Chen, H. S., L. D. Burroughs, and H. L. Tolman, 2003: Ocean surface waves. NWS/NCEP Technical Procedures Bulletin 494.
- [4] Koller, D., Friedman, N. Probabilistic Graphical Models: Principles and Techniques (Adaptive Computation and Machine Learning). The MIT Press, August 2009.
- [5] LeRay, Phillipe. BNT Structure Learning Package.
<http://banquiseasi.insa-rouen.fr/projects/bnt-slp/>
- [6] Meteorological Assimilation Data Ingest System.
<http://madis.noaa.gov/>
- [7] Murphy, Kevin. Bayes Net Toolbox for Matlab.
<http://people.cs.ubc.ca/~murphyk/Software/BNT/bnt.html>
- [8] National Data Buoy Center. <http://www.ndbc.noaa.gov/>
- [9] Sverdrup, H. U. and W. H. Munk, 1947: Wind, sea and swell: Theory of relations for forecasting. *Publication 601*, Hydrographic Office, U.S. Navy, 50 pp.
- [10] Tolman, H. L., 2003a: Treatment of unresolved islands and ice in wind wave models. *Ocean Modelling*, 5, 219-231.
- [11] Una Ola, "Costa Rica Surfing Reports," November 4, 2009. [Online]. Available: <http://www.unaola.com/costa-rica-surfing/surf-reports/>. [Accessed Nov 4, 2009]
- [12] Wikipedia: Relative Humidity.
http://en.wikipedia.org/wiki/Relative_humidity