Spatiotemporal Bayesian Model for Predicting the Occurrence of Surface Objects

Yang Cai[1], Karl Fu[1], Xavier Boutonnier[1], Rafael Franco[1], Daniel Chung[1],
Richard Stumpf[2], Timothy Wynne[2], Mitchell Tomlison[2]
[1] Carnegie Mellon University, ycai@cmu.edu
[2] NOAA, richard.stumpf@noaa.gov

***Abstract* - In this paper, we present a spatiotemporal Bayesian model for predicting the occurrence of surface objects. In the model, we divide a space into a set of two-dimensional lots where each lot has a set of states and historical data. The model is tested with the Karenia brevis cell counts and other data from the coast of Florida for eight years, from 1997 to 2002. From 5,000 random samples, we reach an about 87.5% of chance to correctly predict the presence of a toxic cell count level at a specified location. However, the false alarm rate is 51.95%, which means that a non-toxic cell count level is incorrectly classified to be a toxic. To further improve the accuracy, we will incorporate the wind and the chlorophyll anomaly data into the existing model.**

I. INTRODUCTION

Remote sensing databases, such as SeaWiFS and MODIS, have been used as means of monitoring the spatiotemporal dynamics of ocean objects, such as harmful algal blooms (HAB) and river plumes in the presence of coastal areas. However, the current HAB computational models are limited as off-line analyses that have not been seamlessly integrated into day-to-day field applications yet. There is a need for advanced computing techniques that could be applied to the automatic detection or tracking of harmful objects, as well as to the physiological status or taxonomic classification of bloom organisms, in near-shore coastal environments, as well as in the open ocean. Evaluating bloom detection techniques has a critical dependence at some level of visual analysis (Tomlinson et al., 2004). To determine chlorophyll or other cardinal property, multiple samples and parametric statistics are appropriate. For nominal properties, such as bloom type, each bloom must be treated as a single unit, regardless of the number of samples for validation. This is a non-parametric problem that cannot use simple pixel statistics, as it requires identifying contiguous blooms.

The spatiotemporal data mining involves object tracking and modeling, which extract patterns from multiple data streams, such as multi-spectrum satellite images, in-situ cell counts, weather data, and qualitative and quantitative models. The problems in predicting the occurrence of surface objects include 1) multi-resolution sensory fusion for satellite images and cell counts, 2) interaction of external forces such as wind and coastal lines with intrinsic properties such as shape and concentration, 3) multi-physics modeling that fuses biological, chemical and fluid dynamics.

Problems in spatiotemporal data-mining arise while classifying data and predicting future events. Current solutions to the problem include Neural Networks and Physical Modeling. Neural Networks arbitrarily fit the data and are able to predict and classify only if conditions are the same as those in the training set. Neural networks also fail at predicting and classifying data if the training set does not cover the range of the input data. Complex and sophisticated physical modeling are able to predict and classify to extreme accuracy if all the parameters are calibrated. The common problem with physical modeling is that there is not enough relevant data to calibrate all the parameters.

In this paper, we present the progress in our project (AIST-QRS-04-3031) "Spatiotemporal Data Mining for Monitoring and Tracking Ocean Objects," sponsored by NASA ESTO-AIST program. The objective of the project is to predicate conditions favorable for an anomalous event to occur where targets have not been observed. We have developed a spatiotemporal Bayesian model for predicting the occurrence of ocean surface objects. Our case study is based on the HAB (harmful algal blooms) database off the coast of Florida. At the current stage, we only use the cell count data and the geographical information and occurrence time as historical data. In the near future, we will use more variables, including the satellite images of chlorophyll and anomaly from NASA and data regarding the HAB (salt concentration, wind, cell count, and time).

II. SPATIOTEMPORAL BAYESIAN MODEL

Naïve Bayesian inference models have been developed for decades. It is so far the most popular statistical method for prediction. The model has following advantages:

- *Incremental:* the more evidences, the more robust prediction.
- *Linear speed: O(N)* process where N is the number of training sets.
- *Recursion:* the model can update new evidence by recursion.

There are many dialects of the Bayesian model. One of the recent additions is the Spatial Bayesian algorithm [20] that has been used in geographical information retrieval and ecological studies. Spatiotemporal Bayesian inference [9] is developed for dipole analysis of neuroimaging data. However, there is no a common algorithm for many different applications. In our case, we have a long period of data but

very sparse in time and space due to the limitation of data acquisition or the noises generated from the data collection process, e.g. the data under a cloud in a satellite image would be missing. Therefore, it is necessary to generalize our problem to be a two-dimensional data mining problem, where the interested objects occur on the surface only. In addition, we focus on how to formulate the spatial and temporal information into the rigid Bayesian model and also draw the links between the environmental factors and the internal variables.

2.1 Spatial Bayesian Prediction Definitions

Given a lot with $x \in \{0,1,2,...,m\}$, $y \in \{0,1,2,...,n\}$ as shown in Figure 1, we can divide a 2-D space into a grid of lots.

(0,0)	(1,0)	(2,0)	…	(m,0)
(0,1)	(1,1)	(2,1)	…	(m,1)
(0,2)	(1,2)	(2,2)	…	(m,2)
…	…	…	…	…
(0,n)	(1,n)	(2,n)	…	(m,n)

Fig.1. Definition of the 2-D lots

Let the upper left corner of lot *(0,0)* and the lower right corner of the lot (m, n) enclose all of the data. Assume *j* is one of 2 possible states (toxic or non-toxic). Let $v_j^{x,y} \in V$ represent a possible state of lot *(x,y).* The $v(x,y,t)$ will be the predicted state of lot *(x,y)* based on the largest probability of all possible states of $v_j^{x,y} \in V$. The spatiotemporal Bayesian prediction equation is described as below:

$$v(x,y,t) = \arg\max_{\forall j} P(v_j)P(t \mid v_j)P(x \mid v_j)P(y \mid v_j) \qquad (1)$$

In many cases, there are a lot of evidences such as wind, salinity, and other variables in a grid. Assume is evidences e_k *(k = 1,...i).* The prediction equation (1) is modified to

$$v(x,y,t) = \arg\max_{\forall j} P(v_j)P(t \mid v_j)P(x \mid v_j)P(y \mid v_j)\prod_{k=1}^{i} P(e_k \mid v_j) \qquad (2)$$

3.2 Spatial Prediction Confidence

A Spatiotemporal Bayesian Prediction of the lot *(x,y)* at time t is $v(x,y,t)$. However, after the prediction, a confidence level of the prediction may be needed for purposes such as visualization.

Let the confidence value of the prediction $v(x,y,t)$ of lot *(x,y)* at time t be denoted as $C(x,y,t)$. The confidence level of a prediction is the Bayesian probability of state *j* in which the argument in the equation (2) is maxed. So $C(x,y,t)$ is given by the equation

$$C(x,y,t) = \frac{P(v_j)P(t \mid v_j)P(x \mid v_j)P(y \mid v_j)}{P(x,y,t)} \qquad (3)$$

The equation above is derived using eq. (2) divided by the independent probability of all the evidence presented. Figure 2 shows an example of the visualization result of the confidence probability in form of an iso-surface, where all the points on the surface have P= 0.5. The iso-surface illustrates the spatiotemporal movement pattern of the interested object at a defined confidence level.

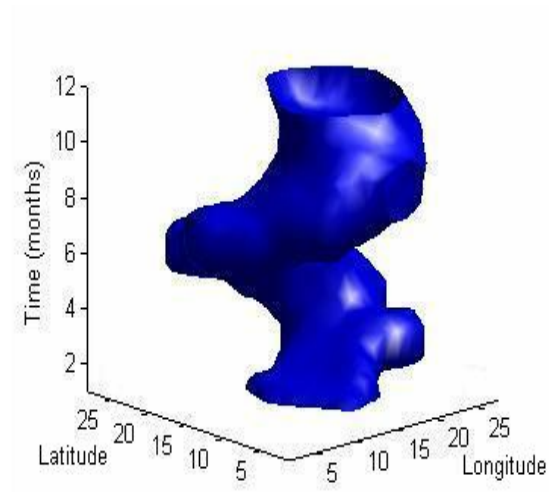

Fig.2. An iso-surface of the probability of the predicted spatiotemporal dynamics

III. PREDICTION OF THE OCCURANCE OF HAB

The data provided for the HAB case study by NASA and NOAA are eight years of satellite images, salt concentration, wind, ocean current, and cell count, from 1975 to 2002. The data come from different stations on the ocean, so its latitude and longitude are the same at one station.

Initially we have 12,616 sets of training data. Each set is guaranteed to have to have a date, latitude, longitude, and cell count information which are guaranteed to be present. In our test, we randomly removed 3,000 samples, one at a time, from the training set, and tested it. By using our Spatiotemporal Bayesian Prediction model, we calculate the probability of HAB being toxic (denoted T, cell count ≥

5,000) and the probability that it is not toxic (denoted N, cell count < 5,000). After each test, we put the sample back into the database.

For example, after removing an entry with time t, latitude x, longitude y, and cell count C, we find the probability of toxic and non-toxic using $v(x, y, t) = T$ if,

$P(C \geq 5{,}000)P(t \mid C \geq 5{,}000) > P(C < 5{,}000)P(t \mid C < 5{,}000)$,

otherwise $v(x, y, t) = N$.

When calculating the probability of t given an event, a 15 day period span on both sides of t is also incorporated into the prediction to avoid sparse data.

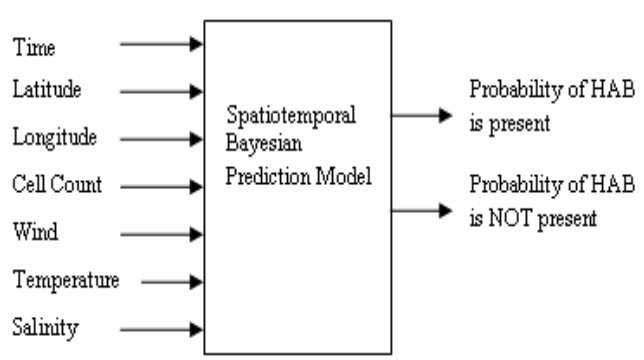

Fig. 2. Inputs and outputs of the Spatiotemporal Bayesian Model (We only use the first four variables at this moment)

In our case, we have sparse data that yield accuracy problems, i.e. N_c is small. Therefore, we used follow formula to estimate $P(e_k \mid v_j) = \frac{N_c + mp}{N_v + m}$, where N_c *is* the number of training instances with evidence e_k and state v_j and m is the constant to enlarge the sample size, p is the prior estimate of the probability such that p = 1/r, where r is number of values that e_k can take, if we assume the prior is uniform.

The inputs into the model are time, latitude, and longitude. The output is the probability the HAB is present, and the probability the HAB is not present. The two probabilities are compared to each other to determine the actual prediction. For example, if probability of HAB present is greater than the probability of HAB not present, then the prediction result is present.

IV. Analysis of the Results

Using the method described above, of all the 5,000 random samples that had toxic cell count level, 87.0% of the predictions of the toxic level were accurate; and of all the 5,000 samples that had a non-toxic cell count level, 51.9% were classified as the toxic level. Therefore, the false alarm rate is rather high at this moment. Table I and II show the detailed results from two trials for each case.

TABLE I. Correct prediction of toxic

Trial	Correct/Total samples	Accuracy
1	1695/1938	87.5%
2	1687/1927	87.5%
Average	-	87.5%

TABLE II. False alarm rate (non-toxic as toxic)

Trial	Correct/Total samples	Accuracy
1	1598/3062	52.2%
2	1589/3073	51.7%
Average	-	51.95%

From the preliminary results, it is obvious that we need more multimodal variables to make a more accurate prediction. At the current stage, we mainly use the spatiotemporal cell-count historical data to make the prediction. We are currently working on incorporating the anomaly and chlorophyll channel data from SeaWiFS satellite images into the prediction model. As the reference [19] shows, there is a linear correlation between the chlorophyll anomaly in the images and Karenia brevis blooms. Therefore, it would significantly increase the accuracy in prediction. The movement of the HABs is also correlated with the wind data. We are incorporating the wind database into our prediction model. Furthermore, we would explore other potential correlated variables such as temperature and salinity.

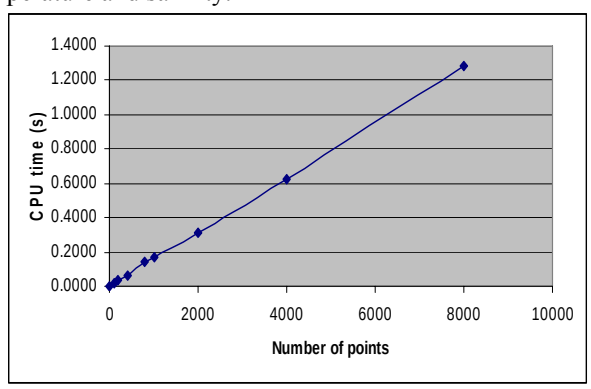

Fig. 3. Number of data point vs. CPU time

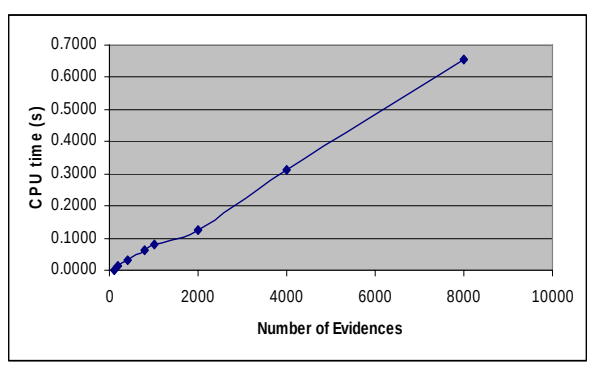

Fig. 4. Number of evidences vs. CPU time

We investigated the computational performance of the model and found the near linear relationship between the CPU time and number of data points or number of pieces of evidences. This proves that the model is 'cheap' in terms of computing resources.

V. CONCLUSIONS

In this paper, we present a Spatiotemporal Bayesian model for predicting the occurrence of surface objects. In the model, we divide a space into a set of two-dimensional lots where each lot has a set of states and historical data.

The model is tested with Karenia Brevis cell counts and other data from Florida coast for 40 years. Using this method, of all the random samples that had toxic cell count level, 87.0% of the predictions were accurate; However, the false alarm rate is 51.95%, which is rather high.

From our computational performance tests, we found that there is a near linear relationship between the CPU time and number of data points or number of evidences. This proves that the model is 'cheap' in terms of computing resources.

Further investigation includes incorporating chlorophyll anomaly data from satellite images, along with wind data and recursive learning.

ACKNOWLEDGEMENT

This study is supported by NASA ESTO grant AIST-QRS-04-3031. We are indebted to our collaborators in NOAA, The authors appreciate the comments and suggestions from Karen Meo, Kai-Dee Chu, Steven Smith, Gene Carl Feldman and James Acker from NASA. Also, many thanks to Professors Christos Faloulus and Mel Siegel of Carnegie Mellon University for their input.

REFERENCES

[1] Bretthorst, G. Larry, 1988, *Bayesian Spectrum Analysis and Parameter Estimation* in Lecture Notes in Statistics, 48, Springer-Verlag, New York, New York;

[2] Cai, Y. and S. Chung, R. Stumpf, T. Wynne, M. Tomlinson, et al, Spatial Interaction Model for Monitoring Harmful Algae Blooms, proceedings of NASA ESTC-05 Conference, Washington DC, 2005

[3] Cai, Y. and Y. Joen, Cellular Metaphor Augmented Spatiotemporal Data, International Conference of Multimodal Interfaces, Trento, Italy, 2005

[4] Cai, Y. (editor), Ambient Intelligence for Scientific Discovery, Lecture Notes in Artificial Intelligence, LNAI 3345, Springer, Feb. 2005

[5] Daniel B. Neill and Andrew W. Moore. Anomalous spatial cluster detection. *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, 2005.

[6] Devender Sivia, *Data Analysis: A Bayesian Tutorial*. Oxford: Clarendon Press (1996), ISBN 0-19-851889-7

[7] Framinan, M.B. and O.B. Brown, 1996. Study of the Rio de la Plata turbidity front, Part I: spatial and temporal distribution. Continental Shelf Research, 16(10): 1259-1282.

[8] Jaynes, E.T. (1998) *Probability Theory : The Logic of Science*.

[9] Jun S.C., George J.S., Pare-Blagoev J., Plis S.M., Ranken D.M., Schmidt, D.M., Wood, C.C. Spatiotemporal Bayesian inference dipole analysis for MEG neuroimaging data. Neuroimage. 2005 Oct 15;28(1):84-98. Epub 2005 Jul 15.

[10] Kulldorff,M. A spatial scan statistics. Communications in Statistics: Theory and Methods, 26(6), 1481-1496, 1997

[11] Maheshkumar R. Sabhnani, Daniel B. Neill, Andrew W. Moore, Fu-Chiang Tsui, Michael M. Wagner, and Jeremy U. Espino. Detecting anomalous patterns in pharmacy retail data. *Proceedings of the KDD 2005 Workshop on Data Mining Methods for Anomaly Detection*, 2005.

[12] McCallum, A. and Nigam K. "A Comparison of Event Models for Naive Bayes Text Classification". In AAAI/ICML-98 Workshop on Learning for Text Categorization, pp. 41-48. Technical Report WS-98-05. AAAI Press. 1998. *(available online: PDF).*

[13] Menzies, T. and Y. Hu, Data Mining For Very Busy People. IEEE Computer, October 2003, pgs. 18-25.

[14] Neill, D., Moore, A.W. Sabhnani, and Daniel, K. Detection of emerging space-time clusters. *Proceedings of the 11th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 218-227, 2005.

[15] Salisbury, J.E., J.W. Campbell, L.D. Meeker, C. Vorosmarty, 2001. Ocean color and river data reveal fluvial influence in coastal waters. EOS, Transactions, American Geophysical Union, 82(20): 221, 226, 227.

[16] Stumpf, R.P., 1988, Sediment transport in Chesapeake Bay during floods: analysis using satellite and surface observations: Journal of Coastal Research, v. 4 p. 1-15.

[17] Stumpf, R.P. and P. Goldschmidt, 1992, Remote sensing of suspended sediment discharge in the western Gulf of Maine, April 1987 100-year flood. Journal of Coastal Research, v. 8, p. 218-225.

[18] Tomlinson, M.C., R.P. Stumpf, V. Ransibrahmanakul, E.W. Truby, G.J. Kirkpatrick, B.A. Pederson, G.A. Vargo, C. A. Heil., 2004. Evaluation of the use of SeaWiFS imagery for detecting Karenia brevis harmful algal blooms in the eastern Gulf of Mexico. Remote Sensing of Environment, v. 91, pp. 293-303.

[19] Walker, A.R., Pham, B. and Moody, M. Spatial Bayesian Learning Algorithms for Geographic Information Retrieval, Proceedings of the 13th annual ACM international workshop on Geographic information systems, Bremen, Germany, 2005