

## **Tactical Predictions of Shoreline Oiling Probability via Machine Learning Models and Satellite-Derived Surface Oil Analysis Products**

Zachary Nixon

Research Planning, Inc., 1121 Park St., Columbia, SC 29201 USA

### **ABSTRACT 300251:**

For significant oil spills in remote areas with complex shoreline geometry, apportioning Shoreline Cleanup Assessment Technique (SCAT) survey effort is a complicated and difficult task. Aerial surveys are often used to select shoreline areas for ground survey after an initial prioritization based upon anecdotal reports or trajectory models, but aerial observers may have difficulty locating cryptic surface shoreline oiling in vegetated or other complex environments. In dynamic beach environments, stranded shoreline oiling may be rapidly buried, making aerial observation difficult. A machine learning-based model is presented for estimating shoreline oiling probabilities via satellite-derived surface oil analysis products, wind summary data, and shoreline habitat type and geometry data. These inputs are increasingly available at spatial and temporal scales sufficient for tactical use, enabling model predictions to be generated within hours after satellite remote sensing products are available.

The model was constructed using SCAT data from the *Deepwater Horizon* oil spill, satellite-derived surface oil analysis products generated during the spill by NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) using a variety of satellite platforms of opportunity, and available shoreline geometry, character, and other preexisting data. The model involves the generation of set of spatial indices of relative over-water proximity of surface oil slicks based upon the satellite-derived analysis products. The model then uses boosted regression trees (BRT), a flexible and relatively recently developed modeling methodology, to generate calibrated estimates of probability of subsequent shoreline oiling based upon these indices, wind climatological data over the time period of interest, and other shoreline data. The model can be implemented via data preparation in any Geographic Information System (GIS) software coupled with the open-source statistical computing language, R. The model is entirely probabilistic and makes no attempt to reproduce the physics of oil moving through the environment, as do trajectory models. It is best used in concert with such models to make estimates at different spatial scales, or when time and data requirements make implementation of fine-scale trajectory modeling impractical for tactical use. The details of model development implementation and assessments of model performance and limitations are presented.

### **INTRODUCTION:**

During the Mississippi Canyon 252 (MC252) /*Deepwater Horizon* spill, apportioning Shoreline Cleanup Assessment Technique (SCAT) survey effort was a complicated and difficult task. Aerial surveys were often used to select shoreline areas for ground survey after an initial prioritization based upon anecdotal reports or trajectory modeling, but aerial observers occasionally had difficulty locating cryptic surface shoreline oiling in vegetated or other complex

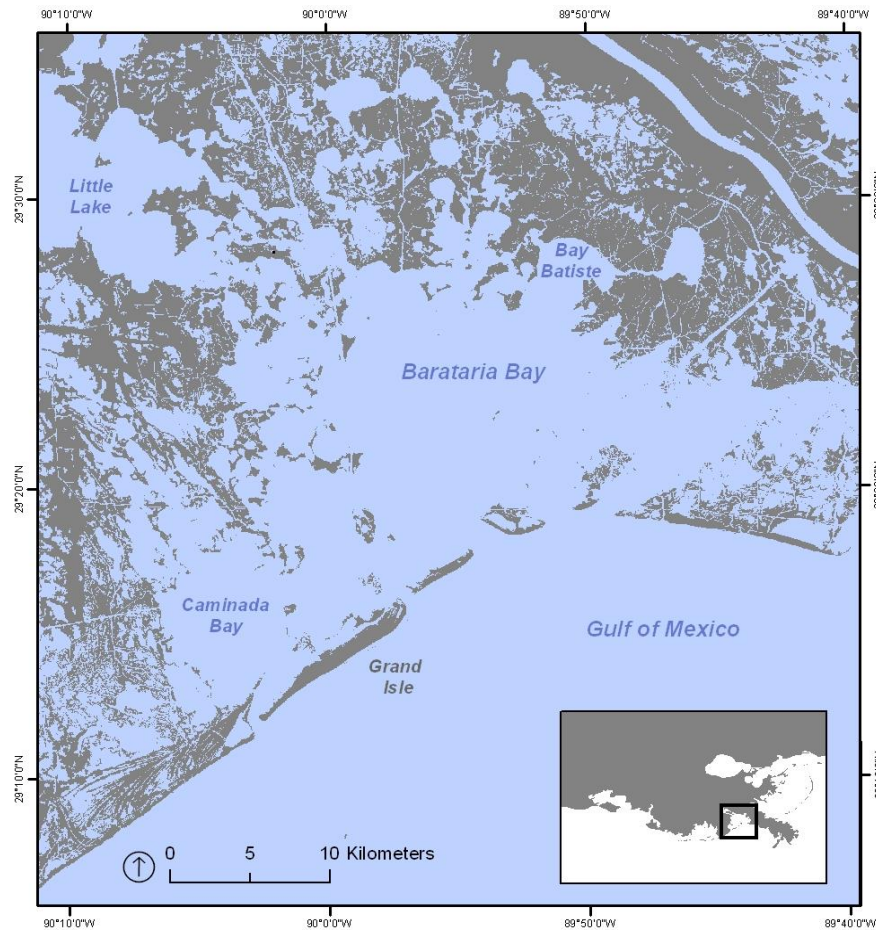
environments. In dynamic beach environments, stranded shoreline oiling is rapidly buried, making aerial observation difficult. Similarly, trajectory modeling is a well established and effective method for estimating potential shoreline oil exposure, but such models may be difficult to use tactically in areas with poorly understood bathymetry and hydrodynamic conditions. This was true during the MC252 spill. Further, the long spill duration led to continual re-oiling at irregular intervals and previously surveyed shorelines would often need to be resurveyed.

Recent advances in machine-learning and statistical modeling methods make it possible to bring more powerful non-physical modeling methods to bear quickly and efficiently. An iterative machine learning-based modeling procedure is presented for estimating relative shoreline oiling probabilities via satellite-derived surface oil analysis products, wind summary data, and shoreline habitat type and geometry data for use at tactical timescales. The model involves the generation of multiple indices of shoreline oil loading and then uses boosted regression trees (BRT), a flexible and relatively recently developed modeling methodology, to generate calibrated estimates of relative probability of shoreline oiling based upon these indices, wind climatological data over the time period of interest, and other preexisting or rapidly available shoreline data. All model inputs are expected available at spatial and temporal scales sufficient for tactical use, enabling model predictions to be generated within hours after satellite remote sensing products are available and in time to prioritize locations for immediate survey.

## **METHODS:**

The model was constructed primarily from satellite-derived surface oil analysis composite daily products generated during the Mississippi Canyon 252 (MC252) /*Deepwater Horizon* oil spill by NOAA's National Environmental Satellite, Data, and Information Service (NESDIS) using a variety of satellite platforms of opportunity (NOAA, 2013), as well as available preexisting shoreline geometry and general morphology. These experimental NESDIS Marine Pollution Surveillance Daily Composite Products form the core of the approach. These products depict the outline of a visible anomaly interpreted as oil by an analyst from a variety of satellite sources including COSMO SKYMED 2, ENVISTAT ASAR NASA MODIS AQUA, and RADARSAT-1. These products were generally available operationally for the first time at a large spill during MC252. The products are experimental and should be used with caution, but represent an important advance in terms of synoptic or near-synoptic, semi-quantitative data depicting on-water oil location at tactically relevant operational timescales.

The model was trained using survey-derived, shoreline surface oiling presence-absence data from SCAT data collected during the spill (Michel et al., 2013) but all model inputs are expected to be available either before future spills or as near real-time products during any future spills of significance. This process was performed for a subset of the MC252 impact area in Louisiana consisting of Barataria Bay and associated barrier islands (Figure 1). This area was selected to provide a good range of oiling conditions and shoreline geometries in a complex environment that was difficult to prioritize for survey during the MC252 incident.



**Figure 1.** Study area in Barataria Bay.

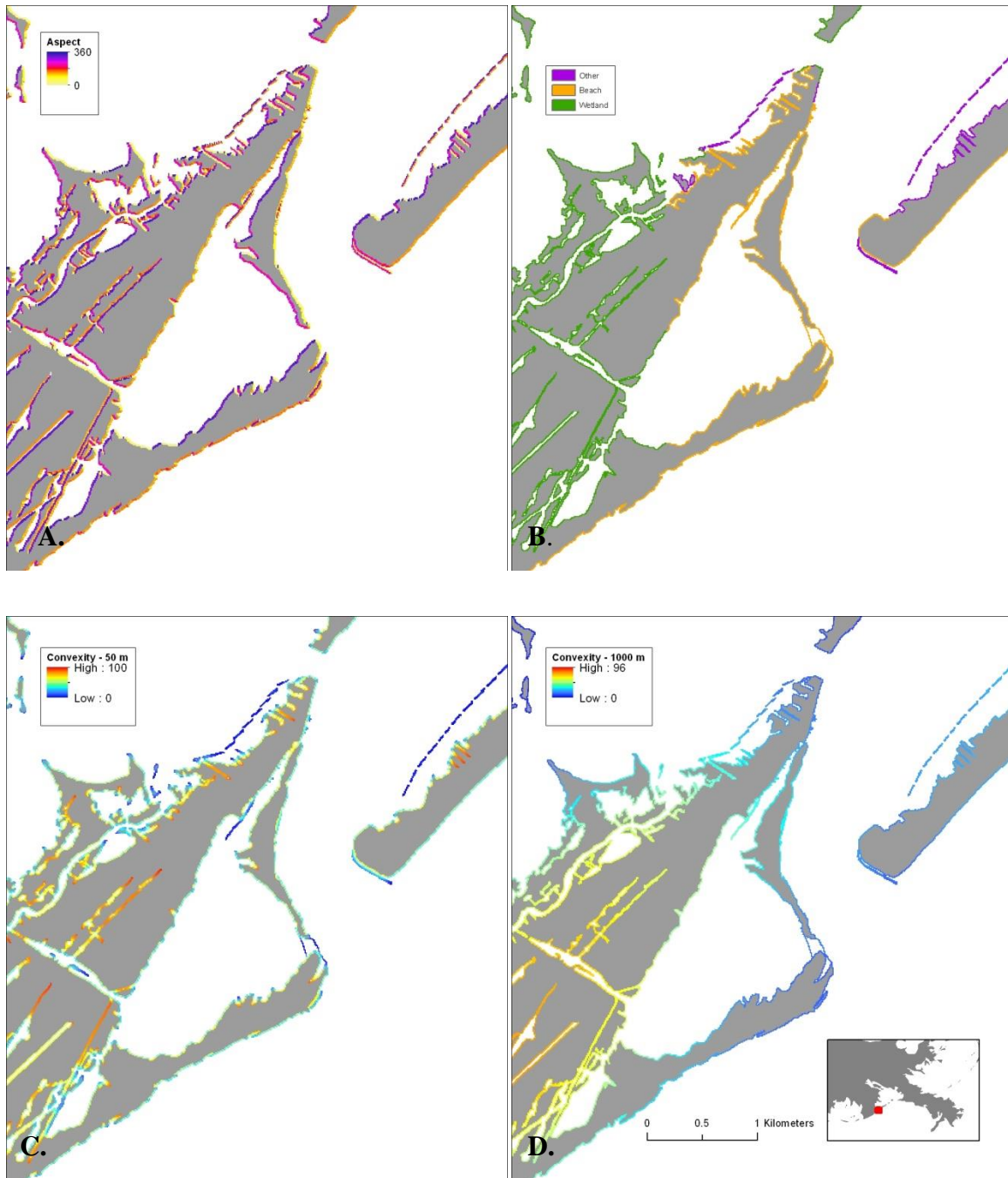
### Predictor Variables

First, a common raster grid was created for the study area using a 10-meter (m) by 10-meter cell size and a series of raster grids representing shorelines and various shoreline geometry variables were generated. A land-water raster and accompanying shoreline raster were then created using this standard grid. We then computed shoreline aspect for all shoreline cells in this 10-m raster by converting land-water to raster with land having a value of 1, water having a value of -1. Aspect was calculated for this raster in ArcGIS 10.2 (ESRI, 2013). Indices of shoreline convexity/concavity may influence the wave and current energy incident upon a shoreline and, thus, may affect the deposition and persistence of stranded oil. A continuous index of concavity/convexity was calculated for all shoreline cells in a 10-m raster. This index was calculated, for each cell, as the arithmetic mean of all cells within a given radius in a land/water raster grid wherein cells were coded as 0 for water and 1 for land. This yielded a unitless index ranging from near zero (extremely convex) to near one (extremely concave). This index was calculated using multiple radii: 50 m and 250 m and 1,000 m. Lastly, a simplified habitat classification was generated based upon the Environmental Sensitivity Index (ESI) shoreline codes assigned to shoreline segments by SCAT teams wherein a shoreline location was classified

as beach, wetland, or other. Figure 2 depicts the rasters of shoreline aspect, the 50-m and 1,000-m convexity indices and habitat classification for a small subset of the study area.

The next step was the creation of a set of spatial indices of potential initial oil loading of shoreline based upon estimated on-water oil locations as described in the NESDIS satellite-derived analysis products, together with shoreline geometry and daily wind directions. The index does not account for, nor attempt to reproduce the physical mechanics of oil transport at or near shorelines. Instead, it is based upon the assumptions that shoreline locations within or adjacent to estimated on-water oil footprints and oriented normal to the angle of wind direction are more likely to accumulate and retain surface oil.

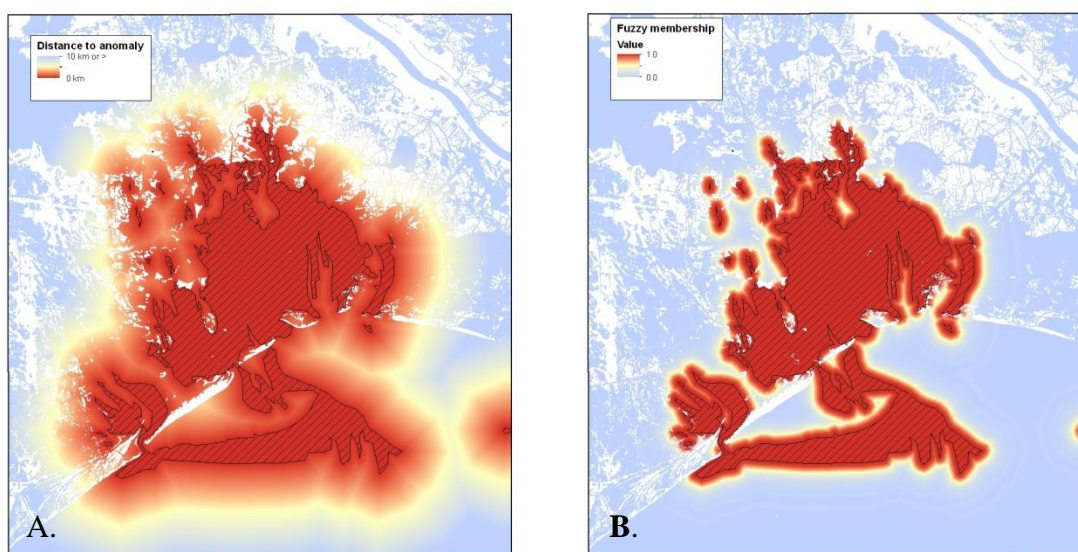
The input data consisted of individual daily composite on-water footprint polygons from the NESDIS satellite-derived analysis products. Because the extent of each of these daily polygons is somewhat imprecise and uncertain, and because the polygons represent an interpreted on-water oil footprint at a specific time which is, of course, anticipated to move over the water surface before the next time of the next NESDIS daily composite, it was required to capture the inherent fuzziness of the boundaries of these polygons in space and time. As such, each polygon was converted to a binary raster using the standard grid, and an over-water cost distance to this polygon was computed in ArcGIS 10.2 using the land-water raster as a cost surface. This cost surface was then used as input to a fuzzy membership function in ArcGIS 10.2. The membership function was a decreasing fuzzy small function which fits a decreasing sigmoid curve through the over-water distance values such that a distance of 0 (inside the daily NESDIS polygon) is redefined as 1, an overwater distance of 1,000 m is defined as 0.5, and an overwater distance of 2,000 m is defined as 0.1. See Lou and Dimitrakopoulos (2003) or Tsoukalas and Uhrig (1997) for more information about fuzzy membership functions and their application in the geosciences. The parameters used to generate this fuzzy membership function are somewhat arbitrary, but are based upon a qualitative assessment of NESDIS polygons accuracy and estimates of average on-water slick movement over single day intervals. Further investigation into the relationship between spatial accuracy and movement potential and how these factors might be better translated into a fuzzy function for transforming over-water distance will be a fruitful area of research as these types of products are increasingly used operationally.



**Figure 2.** Shoreline aspect (A), habitat (B) and 50 meter (C) and 1000 (D) meter convexity index rasters for a subset of the study area near the south end of Grand Isle.

Figure 3 depicts the NESDIS anomaly, the overwater distance raster, and the fuzzy membership raster for June 10, 2010.





**Figure 3.** NESDIS anomaly polygon and the resulting raster of overwater distance to that polygon (A) and the fuzzy membership raster for (B) June 10, 2010.

Wind data were collected as 6-minute averages of speed and direction from the NOAA National Data Buoy Center (NDBC) GISL1 station at Grand Isle (NOAA, 2012) along the southern boundary of Barataria Bay. Daily averages of wind speed and percentages of all observations were computed in 8 directional bins each centered about cardinal directional headings representing  $45^\circ$  intervals for each day. The daily index ( $i_d$ ) was calculated for each shoreline cell for each available daily NESDIS anomaly composite ( $d$ ) as such:

$$i_d = \sum_{a=0^\circ}^{315^\circ} \left( \frac{\cos(a - n) + 1}{2} * p_{ad} \right) (f_d)(s_d)$$

where  $d$  is a given date of a NESDIS anomaly composite,  $a$  is one of eight cardinal wind direction radials ( $0^\circ - 315^\circ$ ),  $n$  is the angle normal to the shoreline, and  $p_{ad}$  is the percentage of time the wind blew from within a  $45^\circ$  cone centered about angle  $a$  on day  $d$ ,  $f_d$  is the output of the fuzzy membership function applied to the distance to the NESDIS anomaly composite on date  $d$ , and  $s_d$  is the average wind speed in meters per second on day  $d$ .

For any series of individual loading index values ( $i_d$ ), we than may construct a normalized composite index ( $i_c$ ) of shoreline loading for each shoreline cell as such:

$$i_c = \sum_{d=1}^d (i_d) / \max \left( \sum_{d=1}^d (i_d) \right)$$

where max is the maximum global value for any cell for all summed individual loading index values. A single normalized composite loading index raster was generated for all NESDIS dates for initial model parameterization, as well as other date ranges as described below.

### Model Parameterization

Parameterization of the model began by generating 10,000 randomly distributed points along all shorelines sampled by the SCAT teams between the beginning of the spill and the end of 2011. Then, 5,000 random points along shorelines identified as oiled, and 5,000 random points along shorelines identified as unoiled were generated. Both samples were generated by using a spatially balanced Generalized Random Tessellation Sample (GRTS) procedure as implemented in ArcGIS 10.2 (Steven and Olsen, 2004; Theobald et al., 2007) to ensure spatially well-distributed random samples.

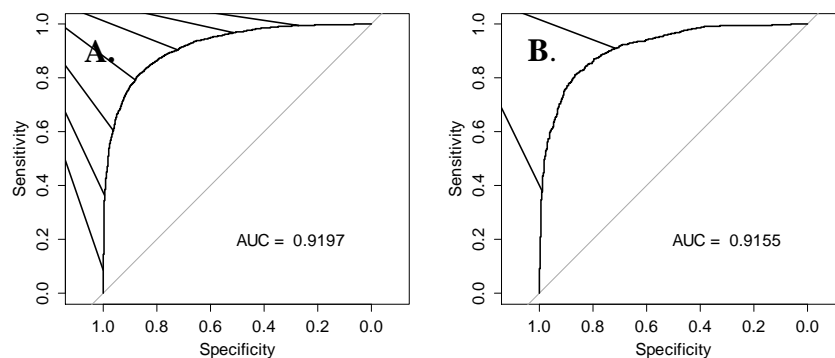
Rasters describing shoreline convexity, shoreline morphology, and the normalized composite shoreline oil-loading index at all points were generated and these values were used to construct the model via an ensemble of boosted regression tree (BRT) classifiers. BRT models attempt to classify observations, here as oiled or unoiled, by repeated recursive binary partitioning of explanatory variable values. The trees are applied sequentially to classify incorrectly classified observations from previous tree models, and the final output is equivalent to a weighted ensemble of hundreds or thousands of tree models. While these models can be more difficult to interpret than more traditional linear models, their performance in classification problems where prediction is the primary goal is generally far superior to many other modeling methods (Elith et al., 2008).

The samples were separated into a 75% training sample and a 25% test sample for more unbiased assessment of model accuracy. The model was constructed using the training data set using the *dismo* package (Hijmans et al., 2013), which in turn makes use of the *gbm* package (Ridgeway, 2013) and the *step.gbm* functions of Elith et al. (2008). The model used a bag fraction of 0.8, a learning rate of 0.01, and a maximum tree complexity of 3. A 10-fold cross validation was used to determine the optimal number of trees for the final model. All other parameters were set to default software settings. All evaluated predictor variables were used as inputs to the BRT model process because of the ability of this method to more rigorously select from competing predictors and reduce the influence of unimportant ones.

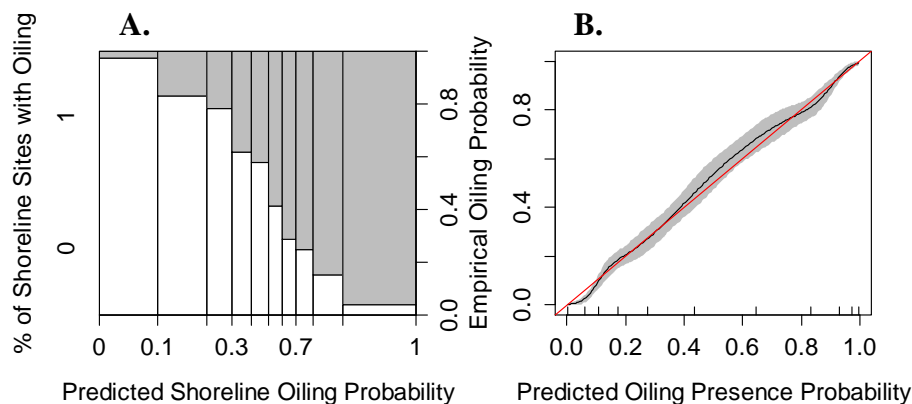
The boosted regression tree algorithm can generally be expected to perform much better than many other competing modeling methods as a binary classifier, but often not as well as an estimator of the true class probability. By definition, logistic models, either single models or averaged ensembles, will more closely approximate the true probability of a class membership. Mease et al. (2007) and Flach and Matsubara (2008) provide more details about this issue. It is possible to calibrate BRT model output to yield class membership probabilities using the scaling method of Platt (1999) as suggested in Niculescu-Mizil and Caruana (2005). The model output was evaluated with spine plots and calibration plots, and it was determined that the models used here were very well calibrated and that Platt scaling was not required here was not required for the model as used here.

## RESULTS:

The performance of the shoreline-oiling model constructed as described above was evaluated by computing the Area Under the Curve (AUC) statistic of Receiver Operating Characteristic (ROC) curves (Robin et al, 2011) for both the training data and the independent test data. While there exist some questions regarding the optimality of AUC as a model performance metric (Jiménez-Valverde, 2012), it is still the most applied measure of accuracy for binary classification models. The performance of the model was very good, with nearly identical AUC values for both training and testing data. Only a slight decline in AUC for the independent test set was observed, and both AUC values were above 0.91 (Figure 5). Spine plot and calibration plots indicated that the model was well calibrated, and that Platt scaling was not required to yield relative probability (Figure 6.)



**Figure 5.** ROC plots and Area Under the Curve (AUC) values for training (A.) and independent test samples (B.)

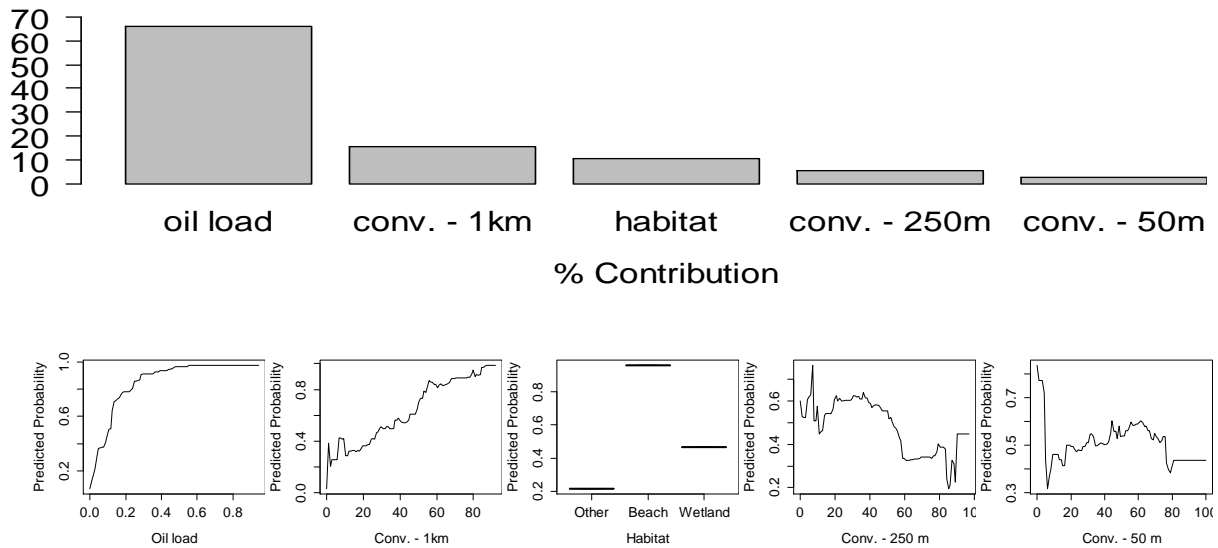


**Figure 6.** Spine plot (A) and calibration plot (B) of predicted and actual empirical probabilities from independent test data.

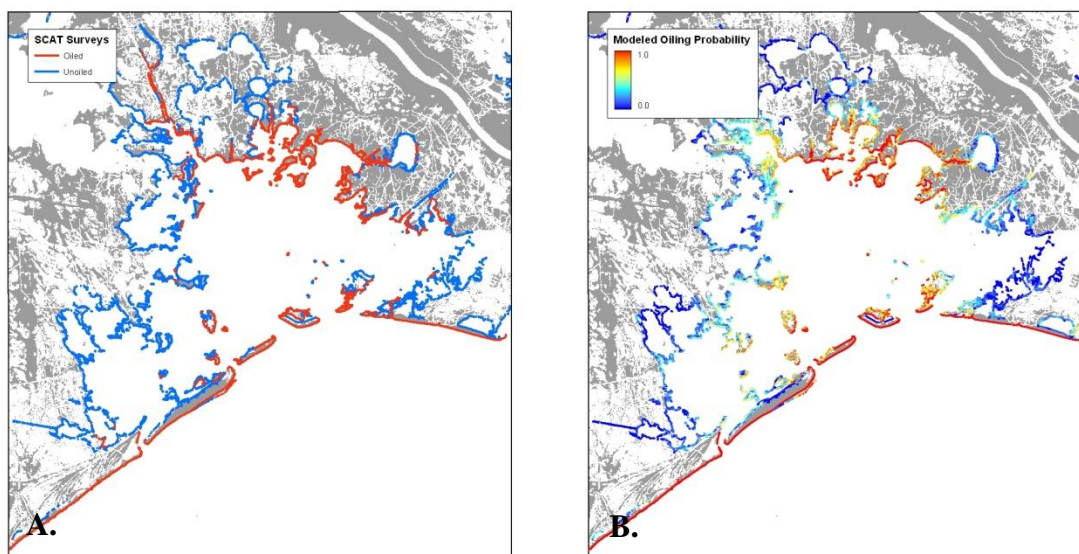
The relative importance of each variable in the model was evaluated by computing the reduction in deviance attributable to each variable in predicting the gradient of the following iteration (Figure 7). This is presented as the relative influence of each variable in reducing the loss function (Friedman, 2001). The marginal effects of each variable, which depicts the effect of



that variable with other variables integrated out, was plotted (Figure 7). Of the variables included in the model, the normalized composite oil-loading index was responsible for approximately two thirds of the explanatory power of the model, with increased oil-loading index values being associated with increased likelihood of oil presence. All the input variables were contributing predictors with the 1,000-m convexity index and habitat being somewhat more significant. Generally increased convexity at the 1,000-m scale was associated with increased likelihood of oil presence, and beaches were more likely to have been oiled than wetlands or other types of shoreline, given similar oil loading and geometry. The other convexity indices demonstrated a more complex relationship with the resulting likelihood of oiling. A map of all surveyed shorelines by oil presence/absence at the end of 2011 and model output for the same shorelines is depicted in Figure 8.

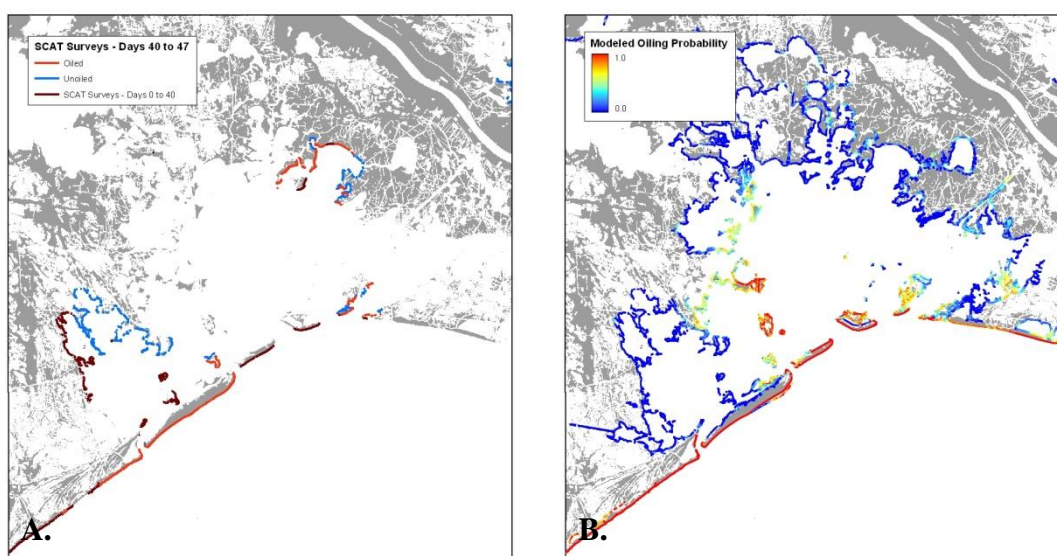


**Figure 7.** Relative contribution (top) and marginal effects plots (bottom) of each variable included in the model.



**Figure 8.** Oil presence/absence at the end of 2011 (A) and model output for the same shorelines (B).

In an operational setting, however, the model would be unlikely to be used in this type of “hindcast” mode. More realistically, it would be used to prioritize as yet un-surveyed shorelines for survey based on recent NESDIS product availability. To simulate the use of the modeling methodology during the MC252 incident, a simulation was conducted running the model at day 40, using only data available at that point, to both create a normalized cumulative oil-loading index and parameterize a new model. The results of that model were compared to actual survey footprints and survey outcomes for the 7 days following. The AUC of this model is as good as the model for the full dataset: 0.919 on an independent test set. Model output is depicted in Figure 9, together with survey outcomes over the next week.



**Figure 9.** Oil presence/absence between 40 and 47 days post-spill (A) and model output for all shorelines based upon normalized composite oiling index values though day 40 post-spill (B).

## DISCUSSION:

It is clear that the model is able to reproduce the relationships between the oil-loading index and SCAT survey outcomes with high fidelity when applied retrospectively. Evaluations of the model performance in the more realistic simulation of an operational setting are more instructive. In this case, fairly large survey effort was apportioned to the bay-rim marsh shorelines of Caminda Bay during the period following simulated tactical use, though model output indicates that effort would have been more productively spent further east along back-barrier marsh shorelines, and the marsh complexes dividing Caminada and Barataria Bays. Obviously, this model output was not available at the time of actual survey, and there were a

variety of decision-making and prioritization criteria driving survey prioritization during the actual spill other than likelihood of oiling. Nevertheless, we feel that the timely availability of shoreline oiling probability based on rigorous analysis of quantitative data has the potential to improve survey prioritization. In some cases, this improvement may be dramatic where manual prioritization by experts is particularly difficult.

It is worthwhile to consider the applicability of this method to future incidents in other geographic settings. The most critical inputs to the model are the satellite-derived surface oil polygons and adjacent shoreline geometry and character. While the processing and application of data from the satellite platforms used here was novel for the MC-252 incident, similar raw data should be, in theory, be available from these platforms for any future incident in any location globally. Detailed shoreline position and character data are also critical predictive inputs, but these data are less likely to be immediately or reliably available. Indeed, this was the case for MC-252 where current shoreline position information was only available a few months after the spill. However, one of the strengths of the BRT methodology is flexibility in handling missing data. As such, this method can gracefully adapt to availability of increasingly detailed data over time, yielding iteratively improved estimates as absent or coarse-scale shoreline position and morphological data are replaced with data of better accuracy and precision.

Potential improvements to this process are manifold. The nature and parameters of the fuzzy function used to weight distance to NESDIS anomaly boundary were derived based upon expert judgment, but a more thorough investigation into overwater movement rates might lead to development of a more appropriate function or family of functions. Similarly, the NESDIS anomaly products are experimental, and no inclusion of reported uncertainty has yet been attempted. Lastly, inclusion of cost-sensitive model metrics during both model parameterization and model output interpretation could yield a more direct incorporation of the different costs of surveying unoiled shorelines versus failing to survey oiled shorelines.

## **CONCLUSION:**

These results suggest that this or similar modeling methods can fill an important tactical role in planning and conducting shoreline oil surveys, particularly during large spills of long duration, or in areas of complex and extensive geography. These results indicate that these models, which can be implemented in any GIS software coupled with the open source statistical computing language, R, can provide unbiased quantitative guidance for prioritization of survey effort in the situations where it is far from clear how survey effort should be apportioned. While it is imperative that these modeling methods be used concert with and as supplemental to fine-scale physical hydrodynamic trajectory modeling, these models hold promise for filling a critical gap in terms of rapidly and quantitatively making tactical use of remotely sensed data and spatial analysis to bring the right survey and response resources to bear in the right places and times.

**REFERENCES:**

- Elith, J., Leathwick, J. and T. Hastie. 2009. A working guide to boosted regression trees. *Journal of Animal Ecology* 77:802-813.
- ESRI. 2013. ArcGIS Desktop, v10.2. Redlands CA, Environmental Systems Research Institute.
- Flach, P. and E. Matsubara. 2008. On classification, ranking, and probability estimation. In: de Raedt, L. T. Dietterich, L. Getoor, K. Kersting and S. Muggleton (eds.). *Probabilistic, Logical and Relational Learning – A Further Synthesis*. Dagstuhl Seminar Proceedings.
- Friedman, J. 2001. Greedy Function Approximation: A Gradient Boosting Machine. *Annals of Statistics* 29(5):1189-1232.
- Hijmans, R., Phillips, S., Leathwick J., and J. Elith. 2013. *dismo*: Species distribution modeling. R package version 0.8-11. Available online at: <http://CRAN.R-project.org/package=dismo>
- Jiménez-Valverde, A. 2012. Insights into the area under the receiver operating characteristic curve (AUC) as a discrimination measure in species distribution modelling. *Global Ecology and Biogeography* 21:498–507
- Lou, X., and R. Dimitrakopoulos. 2003. Data-driven fuzzy analysis in quantitative mineral assessment: *Computers and Geosciences* 29:3-13.
- Mease, D., Wyner, A., and A. Buja. 2007. Boosted classification trees and class probability/quantile estimation. *Journal of Machine Learning Research* 8:409-439.
- Michel J., Owens, E., Zengel S., Graham, A., Nixon, Z., Allard, T., Holton W., Reimer, P., LaMarche, A., White, M., Rutherford, N., Childs, C., Mauseth, G., Challenger, G., Taylor, E. 2013. Extent and Degree of Shoreline Oiling: *Deepwater Horizon* Oil Spill, Gulf of Mexico, USA. *PLoS ONE* 8(6): e65087.
- National Oceanic and Atmospheric Administration (NOAA) National Data Buoy Center (NDBC). 2012. Climatic Summary Tables. Available online at: <http://www.ndbc.noaa.gov/> (accessed on 11/1/2012).
- National Oceanic and Atmospheric Administration (NOAA) National Environmental Satellite, Data, and Information Service (NESDIS). 2013. Satellite Derived Surface Oil Analysis Products - Deepwater Horizon. Available online at: <http://www.ssd.noaa.gov/PS/MPS/deepwater.html> (accessed on 12/1/2013).
- Niculescu-Mizil, A., and R. Caruana. 2005. Predicting good probabilities with supervised learning. In: L. De Raedt and S. Wrobel, (eds.). *ICML '05: Proceedings of the 22nd International Conference on Machine Learning*. ACM, New York, New York, pp. 625-632.

- Platt, J. 2000. Probabilistic outputs for support vector machines and comparison to regularized likelihood methods . In: A. Smola, P. Bartlett, B. Scholkopf, and D. Schuurmans (eds.): Advances in Large Margin Classifiers. Cambridge, MA.
- R Development Core Team. 2012. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Available online at: <http://www.R-project.org>. (accessed on 11/1/2012).
- Robin, X., N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J. Sanchez, and M. Müller. 2011. pROC: an open-source package for R and S+ to analyze and compare ROC curves. BMC Bioinformatics 12:77.
- Ridgeway, G., with contributions from others. 2013. gbm: Generalized Boosted Regression Models. R package version 2.0-8. Available online at: <http://CRAN.R-project.org/package=gbm>
- Stevens, D., and A. Olsen. 2004. Spatially balanced sampling of natural resources. Journal of the American Statistical Association 99 (465):262-278.
- Theobald, D., Stevens, D., White, D., Urquhart, N., Olsen, A., and J. Norman. 2007. Using GIS to Generate Spatially Balanced Random Survey Designs for Natural Resource Applications. Environmental Management 40: 134-146.
- Tsoukalas, L, and R. Uhrig. 1997. Fuzzy and neural approaches in engineering: New York, John Wiley and Sons, Inc., 587 pp.