ESTIMATION OF SHORELINE PLANT OILING LENGTH: A GEOSPATIAL APPROACH

Prepared for:

Abt Associates
1881 Ninth Street, Suite 201
Boulder, Colorado 80302

Prepared by:

Dr. Pierre Goovaerts
PGeostat, LLC
11487 Highland Hills Drive
Jerome, MI 49249-9588

Date issued: August 30, 2015

1. Introduction

Studies of vegetation death and accelerated marsh erosion following *Deepwater Horizon* have shown that both of these injuries can be related to the degree of oiling on marsh vegetation (e.g., Hester and Willis, 2015; Silliman et al., 2015). Spatial quantification of these injuries thus relies on estimates of how many kilometers of shoreline fell into each of the four stem oiling categories on which these injury determinations were based (0-10%, 10-50%, 50-90%, 90-100%). Vegetation oiling from the *Deepwater Horizon* spill was unevenly distributed across Louisiana marsh environments, and field observations of stem oiling were collected at discrete points (Deepwater Horizon NRDA, 2010). Shoreline exposure categories provide spatially continuous coverage, but the SCAT and Rapid Assessment data on which these categories are primarily based do not include detailed measurements of stem oiling (e.g., Nixon et al., 2015).

One way of quantifying the length of shoreline falling into each stem oiling category is to assume that stem oiling values are evenly distributed in space within shoreline exposure categories, and to calculate the length of each stem oiling category based on proportional assignments within the shoreline exposure framework. However, this approach does not account for clustering of stem oiling data in space. This may be a particular concern for apportionment of stem oiling data that were recorded within the “no observed oil” (NOO) category (i.e. false negatives), since there were thousands of kilometers of shoreline within this category but nonzero stem oiling observations within this category were generally clustered in space.

An alternative is to use a geospatial analysis, which accounts for spatial patterns in stem oiling data, as well as spatial variability in the relationship between stem oiling and shoreline exposure. This document describes a geospatial analysis that was conducted to estimate the expected lengths of mainland herbaceous shoreline in Louisiana falling into each of the four stem oiling categories: 0-10%, 10-50%, 50-90%, 90-100%.

2. Data

The analysis was based on four main types of data:

1. Measurement of the percentage of plant stem oiling from the preassessment dataset (911 “hard” data)[1]
2. Indicators of presence/absence of plant stem oiling from the pre-assessment dataset (185 “soft” data)
3. Oiling exposure category (secondary information) surveyed along the coastline and at 729 of the 911 hard data locations from Nixon et al. (2015). Oiling exposure is classified into

[1] “Hard” data are precise measurements of an attribute, which in this case means a percentage of stem oiling. “Soft” data are imprecise measurements of an attribute, which in this case means we only know whether it was oiled or not; we do not know the percentage of stem oiling (Goovaerts, 1997)

one of the four following categories: No oil observed (NOO), Light oiling, Heavier oiling, and Heavier persistent oiling.

4. Length of shoreline located within 50×50m squares discretizing the Louisiana coastline.

Figure 1 shows where hard and soft data on percentage of plant stem oiling were collected at preassessment (PA) sites. Summary statistics in Table 1 indicate the presence of stem oiling at 39.6% of hard data sites and only 6% of soft data sites. Similarly, the NOO exposure category is more prevalent at soft data sites (81.4%) relative to hard data sites (47.9%); see Table 2. This table also highlights the fact that secondary information on oil exposure was not recorded at a number of sites: 20% for hard data (182) and 36.2% for soft data (67). In other words, the PA survey data extend beyond the boundaries of the shoreline exposure dataset. Oiling exposure category surveyed along the coastline, which was discretized using a 50 m spacing grid, is mapped in Figure 2. The entire grid includes 118,151 nodes under mainland herbaceous marsh and their repartition between four categories of oiling exposure is listed in Table 2 (last column). The percentage of observations in the NOO category is much larger in the shoreline exposure dataset compared to the hard dataset: 85% of the shoreline exposure data are NOO, compared to 47.9% of the PA points, which reflects the preferential sampling of oiled locations during the PA survey.

Table 1. Distribution of hard and soft data between the five classes of percentage of plant stem oiling. Red numbers in parenthesis are percentages of the total number of non-missing data.

% stem Oiling	Hard data (n=911)	Soft data (n=185)
0%	551 (60.4)	174 (94.0)
0-10%	59 (6.5)	11 (6.0)
10-50%	169 (18.6)	
50-90%	80 (8.8)	
> 90%	52 (5.7)	

Table 2. Distribution of hard and soft data between the four categories of oiling exposure. The last column reports the number of shoreline nodes within each oiling exposure category. Red numbers in parenthesis are percentages of the total number of non-missing data.

Oiling exposure category	Hard data (n=911)	Soft data (n=185)	Shoreline (n=118,151)
No Oil Observed (NOO)	349 (47.9)	96 (81.4)	100,418 (85.0)
Light Oiling	172 (23.6)	17 (14.4)	12,234 (10.3)
Heavier Oiling	153 (21.0)	5 (4.2)	4,096 (3.5)
Heavier persistent Oil,	55 (7.5)	0	1,403 (1.2)
Missing	182	67	

3. Methods

The flowchart of Figure 3 illustrates the main steps of the analysis that was validated using a leave-one-out approach[2] at the 729 locations where percentage of stem oiling and oiling exposure categories were both recorded. Following this validation, this process was applied to the entire coastline. The analysis was conducted using the following software: 1) SpaceStat 4.0 (Jacquez *et al.*, 2014) for geographically-weighted regression and variogram modeling, 2) SAS 9.3 (SAS Institute Inc., 2011) for aspatial logistic regression and the creation of ROC curves, 3) SGeMS (Remy *et al.*, 2008) and Gslib (Deutch and Journel, 1998) for cross-variogram modeling and indicator cokriging, and 4) code written by Dr. Goovaerts for data manipulation and computation of expected lengths of shoreline in different categories of plant stem oiling.

3.1. Indicator coding of plant stem oiling data

The analysis started with the coding of each percentage of stem oiling data into a vector of indicators of exceedance of four thresholds z_c = 0, 10, 50, and 90%. Let $\mathbf{u}_\alpha$= (x_α,y_α) be a vector of UTM coordinates representing the geographical location of a stem oiling data, denoted z($\mathbf{u}_\alpha$) for hard data and s($\mathbf{u}_\alpha$) for soft data. The set of four indicators at any hard data location $\mathbf{u}_\alpha$ was then constructed as:

$$i(\boldsymbol{u}_\alpha; z_c) = \begin{cases} 1 & \text{if } z(\boldsymbol{u}_\alpha) > z_c \\ 0 & \text{otherwise} \end{cases} \qquad c = 1, \cdots, 4 \tag{1}$$

For example, the indicator coding of a stem oiling data z($\mathbf{u}_\alpha$)=0.356 will be: i($\mathbf{u}_\alpha$;z_1)=1, i($\mathbf{u}_\alpha$;z_2)=1, i($\mathbf{u}_\alpha$;z_3)=0, and i($\mathbf{u}_\alpha$;z_4)=0 since only the first two thresholds of 0 and 10% are exceeded. At locations of soft data (i.e. presence/absence of oil), an indicator can only be constructed for the first threshold z_1 = 0% since the exact percentage of plant stem oiling is unknown:

$$i(\boldsymbol{u}_\alpha; z_1) = \begin{cases} 1 & \text{if } s(\boldsymbol{u}_\alpha) = 1 \text{ (oil present)} \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

3.2. Predicting probability of plant stem oiling from oiling exposure category

Because the percentage of stem oiling was recorded only at 911 discrete locations, the extrapolation of this sole information to the entire shoreline length is challenging. It is thus beneficial to incorporate the secondary information provided by oiling exposure category since it is available on a continuous basis over large sections of the shoreline. Logistic regression was used to create a predictive model of the likelihood that a threshold z_c = 0, 10, 50, and 90% of stem oiling is exceeded at any location **u** (i.e. sampled location or coastline grid node) on the basis of the oiling exposure category surveyed at that location. In other words, the dependent

[2] A leave-one-out approach means that each observation in the dataset was removed at a time and its value was estimated from the remaining observations (Goovaerts, 1997).

variable is the indicator defined in Eq. (1) while the covariate is the oiling exposure category. Logistic regression was conducted in two different settings.

First, an aspatial logistic regression was performed whereby the predictive model is created using the entire data set (i.e. 729 locations where both hard data and oiling exposure category were recorded) and without accounting for the geographical location of the data. The underlying assumption is that the relationship between oiling exposure category and percentage of plant stem oiling does not change along the Louisiana coastline (assumption of stationarity). This strong assumption was relaxed in a second analysis where logistic regression is conducted at each location **u** using only neighboring data (e.g. *N* closest data or data falling within a window of size *S* centered on **u**). Because each observation is weighted according to its proximity to the center of the window, this type of local regression is known as geographically-weighted regression[3] (Fotheringham *et al.*, 2002; Goovaerts *et al.*, 2015). The implementation of geographically-weighted regression (GWR) requires the selection of two main parameters: 1) the type of weight function, and 2) the type of search strategy for the local regression (i.e. size of the search window *S* or number of neighbors *N*).

The weight w_α assigned to each observation $z(\mathbf{u}_\alpha)$ when conducting GWR at **u** was computed using the following bisquare weight function:

$$w_\alpha = \begin{cases} [1 - \ (d_\alpha/\mathrm{a})^2]^2 & \text{if } d_\alpha < a \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

where d_α is the Euclidean distance between $\mathbf{u}_\alpha$ and **u**, and *a* is the bandwidth or maximum distance for non-zero weights. The bandwidth was here spatially variable and set to the distance between **u** and the most remote observation used in GWR at that location, which allows taking into account the highly variable sampling density along the coastline. Such an “adaptive” weight function was preferred to the use of a fixed bandwidth because the latter tends to generate more extreme coefficients in GWR maps, which directly affects the visual pattern and may contribute to biased interpretation (Cho *et al.*, 2009).

The window size *S* or number of neighbors *N* for GWR had to be large enough to include, for each location **u**, all four levels of oiling exposure category so that logistic regression could be performed. This condition could not be satisfied without using a very large search window, which decreased the efficiency of GWR. To reduce the size of the search window while ensuring the convergence of logistic regression, the “Oiling exposure” variable was incorporated as a continuous instead of an ordinal[4] variable in the predictive model. In other words, the four categories of increasing oiling exposure listed in Table 2 (NOO, lighter oiling, heavier oiling and heavier persistent oiling) were coded numerically as: 1, 2, 3, and 4. The appropriateness of this

[3] Although more sophisticated methods (e.g. Gelfand *et al.*, 2003) exist, their development within a Bayesian framework limits the size of the datasets that can be manipulated in practice. For example, the two case-studies analyzed in Gelfand *et al.* (2003) include 237 and 120 observations, which is three orders of magnitude smaller than the data analyzed in the present study (118,151 grid nodes).

[4] An ordinal variable is similar to a categorical variable. The difference between the two is that there is a clear ordering of the categories; e.g. the “lighter oiling” category clearly falls between the “NOO” and “Heavier oiling” categories.

approach was tested using the following procedure described in Pasta (2009): the ordinal variable is incorporated as both a categorical and a continuous variable in the regression model and if the categorical variable is not statistically significant, treating the ordinal variable as continuous is acceptable. In the present study, the categorical version of the "Oiling exposure" variable was never significant, with *p*-values ranging from 0.099 (z_1, z_2) to 0.724 (z_4) depending on the stem oiling threshold z_c used, which validated its use for GWR. Then, seven different search strategies were investigated to select observations for GWR: using all data within a radius of 3 or 5 km from location **u**, or using the closest 50, 100, 150, 250, or 350 observations. The choice of a search strategy was guided by the results of a validation analysis and Receiver Operating Characteristic (ROC) curves described in Section 3.3 below.

The trade-off cost for the local modeling of relationships between oiling exposure category and percentage of plant stem oiling by geographically-weighted regression is the greater uncertainty attached to GWR probability estimates relative to aspatial regression. This larger uncertainty translates into wider 95% confidence intervals (CI) and probability estimates that might not differ significantly from 0 or 1. For each of the four thresholds $z_{c,}$ probabilities estimated by aspatial and GWR logistic regression were combined according to the following procedure:

a) If the GWR CI is wide enough to include a very low probability threshold *T* and its counterpart (1-*T*), then the GWR estimate was considered unreliable and replaced by the aspatial regression estimate.

b) If the GWR CI includes a very low probability threshold *T* but not its counterpart (1-*T*), then the GWR estimate was considered to be not significantly different from zero and replaced by 0.

c) If the GWR CI includes a very high probability threshold (1-*T*) but not its counterpart *T*, then the GWR estimate was considered to be not significantly different from one and replaced by 1.

d) If none of conditions (a)-(c) was met, the GWR probability estimate was used.

Seven different very small values (0, 0.001, 0.005, 0.0075, 0.01, 0.015, and 0.03) were considered for the probability threshold *T* and as for the search strategy the final choice of *T* was guided by the results of a validation analysis and ROC curves described in Section 3.3 below.

3.3. Validation analysis using Receiver Operating Characteristic (ROC) curve

The accuracy of the predictive model described in Section 3.2 was assessed by comparing at 729 sampled locations $\mathbf{u}_\alpha$ the estimated probability of exceeding a threshold z_c, p*($\mathbf{u}_\alpha$; z_c), with the ground truth i($\mathbf{u}_\alpha$; z_c) defined in Eq. (1). The comparison was based on ROC curves (Swets, 1988) which plot the probability of false positive versus the probability of detection. For each threshold the ROC curve was created by the following procedure:

- Calculate the probability p*($\mathbf{u}_\alpha$;z_c) that a specific threshold z_c is exceeded at each location $\mathbf{u}_\alpha$.
- Classify as oiled (i.e. percentage of plant stem oiling above the threshold z_c) the locations where this probability exceeds a threshold *P* ranging from 0 to 1.
- Compute the probability of detection as the proportion of "true" oiled location data (i.e. percentage of stem oiling larger than the threshold z_c) where p*($\mathbf{u}_\alpha$; z_c) > P. The probability of false positive is calculated as the proportion of locations that are wrongly declared as having elevated percentage of plant stem oiling.

The most efficient algorithm is the one that allows the detection of a larger fraction of oiled locations at the expense of fewer false positives; that is the ROC curve should be as close as possible to the vertical axis; see example in Figure 4. A quantitative measure of the accuracy of the prediction is the relative area under the ROC curve (AUC statistic), which ranges from 0 to 1.

3.4. Updating prior probabilities of plant stem oiling using indicator cokriging

The predictive model described in Section 3.2 does not make full use of all the information available in that: 1) out of a total of 1,096 stem oiling data only 729 data that also included oiling exposure category were used in the regression, and 2) the geographical coordinates of the data are ignored in aspatial regression and indirectly used in GWR since only the separation distance from the center of the search window is accounted for, thereby ignoring any spatial clustering of the data which could bias results. Indicator cokriging (CK) incorporates this missing information into the prediction of the probability of exceeding a stem oiling threshold z_c using the following estimator:

$$p^*_{CK}(\boldsymbol{u};z_c) = \lambda_0 \times \mathrm{p}^*(\boldsymbol{u};z_c) + \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha \times i(\boldsymbol{u}_\alpha;z_c) \quad \text{with } \lambda_0 + \sum_{\alpha=1}^{n(\boldsymbol{u})} \lambda_\alpha = 1 \qquad (4)$$

where the probability estimated using logistic regression, $\mathrm{p}^*(\boldsymbol{u};z_c)$, is combined with indicators of exceedance at n($\mathbf{u}$) neighboring sampled locations, i($\mathbf{u}_\alpha$; z_c), using a set of weights λ which are the solution of a system of linear equations, known as a cokriging system (Goovaerts, 1997). The weights take into account the proximity of the data to the location $\mathbf{u}$ (e.g. closest data receive more weight), the data configuration (e.g. spatially clustered data receive less weight since they provide redundant information), as well as the spatial pattern of the data. Following Deutsch and Journel (1998), Equation (4) can also be interpreted as the statistical updating of the probability derived from the sole oiling exposure category (prior probability p*($\mathbf{u}$;z_c)) using stem oiling data, resulting in a "posterior" probability of exceeding the threshold z_c.

Similar to geographically-weighted regression, a key issue in cokriging is the search strategy to select n($\mathbf{u}$) neighboring plant stem oiling data. The number n($\mathbf{u}$) was arbitrarily set to a maximum of 8 to avoid smoothing the results by averaging the influence of too many observations. Following a discussion with NOAA scientists and interpretation of indicator variograms, the search window was restricted to 1 km for the two highest thresholds (50 and 90%) while it extended to 2 km for the thresholds of 0 and 10% of stem oiling. In other words, the pollution resulting in the highest percentages of stem oiling was viewed as a more local phenomenon than the lowest pollution levels (0 to 10% plant stem oiling). Whenever the search window does not include any stem oiling data, no updating was conducted and the regression estimate was used as final estimate, i.e. prior and posterior probabilities are the same.

For both geographically-weighted regression and cokriging, the selection of neighbors and computation of spatial weight functions were based on Euclidean distances, rather than over-water distances. To explore the sensitivity of the results to that approximation, the set of 729 locations where both the percentage of stem oiling and oiling exposure categories were recorded underwent a change of coordinates following the procedure described by Løland and Høst (2003). This approach, which is based on a multidimensional scaling (MDS) of the 729×729 matrix of over-water distances between each pair of observations, creates a new data configuration (Figure 5) where Euclidean distances between observations approximate the original over-water distances. A geographically-weighted regression of data using the 50 closest neighbors before and after projection by MDS yields probabilities $p^*(\mathbf{u}_\alpha;z_c)$ that were strongly correlated: r=0.977 to 0.989 depending on the threshold z_c. Based on this result and the similarity of variograms computed on both datasets, the additional complexity associated with projecting the entire coastline into a new system of coordinates was deemed not justified.

3.5. Computing expected length of shoreline in each category of plant stem oiling

The expected length of shoreline where the percentage of plant stem oiling exceeds the threshold z_c was computed from the posterior probabilities $p^*_{CK}(\boldsymbol{u};z_c)$ as:

$$L(z_c) = \sum_{j=1}^{N} l_j \times p^*_{CK}(\boldsymbol{u}_j;z_c) \times i_{Prox}(\boldsymbol{u}_j;z_c) \qquad (5)$$

where:

- N=118,151 is the number of grid nodes discretizing the Louisiana coastline under mainland herbaceous marsh,
- l_j is the length of shoreline located within the 50×50 m square centered on node $\boldsymbol{u}_j$, and
- $i_{Prox}(\boldsymbol{u}_j;z_c)$ =1 if the node $\boldsymbol{u}_j$ is within 1 km (thresholds 50 and 90%) or 2 km (thresholds 0 and 10%) of a pre-assessment site, and zero otherwise.

In other words, the expected length of shoreline with plant stem oiling above z_c was computed as the sum of the product of the shoreline lengths within each 50x50m square and the probability that oiling is above z_c within that square. Only locations in the vicinity of a PA site were included in the computation to avoid extrapolating results to sparsely sampled segments of shoreline. The distances of 1 and 2 km correspond to the size of the cokriging search windows described in Section 3.4.

4. Results

Table 3 summarizes the results for the aspatial logistic regression between the oiling exposure categories (4 levels) and the indicator of exceedance of one of the four thresholds z_c of percentage of plant stem oiling. Incorporating the exposure categories as a categorical vs continuous variable has a moderate impact on the estimated probabilities of exceedance, yet the AUC statistic indicates that the predictive power of the two types of model is identical. This result justifies the use of a continuous exposure variable in geographically-weighted regression.

As described in Section 3.2, a sensitivity analysis was conducted to support the choice of a search strategy for GWR and the use of a probability threshold *T* for merging aspatial and geographically-weighted regression estimates. This sensitivity analysis was first performed for z_c = 90% because it is the most critical threshold for quantification of erosion injury from plant stem oiling (e.g., Silliman *et al*., 2015). Table 4 lists the values of the AUC statistic for all 49 combinations of seven search strategies (using all data within a radius of 3 or 5 km from location **u**, or using the closest 50, 100, 150, 250, or 350 observations) and seven probability thresholds *T* (0, 0.001, 0.005, 0.0075, 0.01, 0.015, and 0.03). The threshold *T*=0 means that GWR estimates are never deemed unreliable; hence aspatial regression estimates are never substituted for GWR estimates. As expected, the predictive power of the model decreases as the size of the search window increases and data farther away are used. The AUC statistic exceeds 0.9 only when using the 50 closest observations and the maximum is found for *T*=0.0075, which was therefore used for merging results. Note that the merged probabilities have a greater accuracy (AUC = 0.913) than the result obtained by aspatial regression alone (AUC = 0.769) or GWR alone (AUC = 896). A sensitivity analysis was also conducted for the three other thresholds z_c = 0, 10, and 50% but the search strategy was fixed and set to the 50 nearest neighbors. Table 5 indicates that a threshold *T*=0.0075 is optimum for all four thresholds and results in a greater accuracy (i.e. higher AUC statistic) relative to aspatial regression (Table 3, bottom line) and GWR (Table 4, case *T*=0).

Table 3. Probability of exceeding a percentage of plant stem oiling threshold z_c estimated using aspatial logistic regression with oiling exposure category as a categorical (black number) or continuous (red number) covariate. The predictive power of each model is quantified using the Area Under the ROC Curve (AUC) statistic. The best models are those with an AUC statistic closest to 1.

	Estimated probability of exceeding a stem oiling threshold z_c			
Oiling exposure category	z_c=0%	z_c=10%	z_c=50%	z_c=90%
No Oil Observed (NOO)	0.215 (0.224)	0.160 (0.172)	0.052 (0.060)	0.017 (0.020)
Light Oiling	0.535 (0.485)	0.459 (0.404)	0.192 (0.150)	0.064 (0.053)
Heavier Oiling	0.706 (0.754)	0.688 (0.754)	0.294 (0.330)	0.124 (0.131)
Heavier persistent Oiling	0.945 (0.909)	0.878 (0.909)	0.600 (0.577)	0.290 (0.290)
AUC statistic	0.766 (0.766)	0.774 (0.774)	0.765 (0.765)	0.769 (0.769)

Table 4. Results of sensitivity analysis of how the search strategy and value of threshold *T* for merging GWR and aspatial regression estimates impact the predictive power of the model as measured using the Area Under the Curve (AUC) statistic. Red highlighted value is the final choice used in the modeling, using the 50 nearest data and a probability threshold T=0.0075.

	Probability threshold *T*						
Search strategy	T=0.03	T=0.015	T=0.01	T=0.0075	T=0.005	T=0.001	T=0
Radius = 3 km	0.799	0.829	0.835	0.845	0.842	0.837	0.832
Radius = 5 km	0.805	0.852	0.856	0.855	0.862	0.870	0.866
50 nearest data	0.791	0.849	0.889	**0.913**	0.912	0.911	0.896
100 nearest data	0.796	0.838	0.838	0.856	0.873	0.880	0.867
150 nearest data	0.803	0.831	0.842	0.846	0.857	0.863	0.863
250 nearest data	0.808	0.825	0.831	0.830	0.845	0.847	0.847
350 nearest data	0.818	0.819	0.831	0.829	0.841	0.841	0.841

Table 5. Results of sensitivity analysis of how the value of threshold *T* for merging GWR and aspatial regression estimates impact the predictive power of the model for different stem oiling thresholds z_c. The Area Under the Curve (AUC) statistic is best if close to 1.

	Probability threshold *T*						
Stem Oiling threshold	T=0.03	T=0.015	T=0.01	T=0.0075	T=0.005	T=0.001	T=0
z_c=0%	0.874	0.876	0.876	0.876	0.876	0.876	0.866
z_c=10%	0.874	0.877	0.877	0.877	0.877	0.877	0.868
z_c=50%	0.846	0.860	0.872	0.872	0.874	0.874	0.857
z_c=90%	0.791	0.849	0.889	0.913	0.912	0.911	0.896

Following the sensitivity analysis, aspatial regression and GWR were conducted at 118,151 nodes discretizing the Louisiana coastline under mainland herbaceous marsh. The two sets of probability estimates were then combined, for each stem oiling threshold, using the procedure described in Section 3.2 with *T*=0.0075. The resulting probabilities are referred to as "prior" probabilities as they are based solely on the oiling exposure category recorded at each grid node. Cokriging was applied to update these "prior" probabilities using the additional information provided by indicator coding of hard and soft data (Eqs. 1 and 2). For each threshold, direct and cross indicator variograms were computed and modeled using a combination of one exponential variogram model with range of 600 m and another exponential model with a range of 15 km for thresholds z_c = 0 and 10 %, 10 km for threshold z_c = 50%, and 5 km for threshold z_c = 90%.

The posterior probabilities estimated by cokriging at each grid node were multiplied by the corresponding length of shoreline according to Eq. (5) to compute the expected lengths of shoreline where the percentage of plant stem oiling exceeds each of the four thresholds z_c = 0, 10, 50 and 90%. Table 6 lists the lengths of shoreline falling into classes 0-10%, 10-50%, and 50-90%, which were obtained as the difference L(z_{c+1})-L(z_c).

Table 6. Expected length of Louisiana shoreline under mainland herbaceous marsh falling into different classes of percentage of plant stem oiling (see Equation 5).

% plant stem Oiling	Expected length (km)
0-10%	109
10-50%	225
50-90%	628
> 90%	199
Total	1,161

5. References

- Cho, S., Lambert, D. M., Kim, S. G., and S. Jung. 2009. Extreme coefficients in geographically weighted regression and their effects on mapping. *GIScience & Remote Sensing, 46*(3), 273-288.
- Deepwater Horizon NRDA, 2010. Deepwater Horizon/MC252/BP Shoreline/Vegetation NRDA Pre-assessment Data Collection Plan. July 12, 2010
- Deutsch, C.V. and A.G. Journel. 1998. GSLIB*: Geostatistical Software Library and User's Guide, 2nd Ed.* Oxford University Press, New York, NY.
- Fotheringham, A. S., Brunsdon, C., & Charlton, M. 2002. *Geographically weighted regression: The analysis of spatially varying relationships* John Wiley & Sons.
- Gelfand, A.E., Kim, H.-J., Sirmans, C.F. and S. Barnejee. 2003. Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, **98**: 387–396.
- Goovaerts, P. 1997. *Geostatistics for Natural Resources Evaluation*. Oxford University Press, New-York, NY.
- Goovaerts, P., Xiao, H., Adunlin, G., Ali, A., Tan, F., Gwede, C.K., and Y. Huang. 2015. Geographically-weighted regression analysis of percentage of late-stage prostate cancer diagnosis in Florida. *Applied Geography*, **62**: 191-200.

- Hester, M. W., J. M. Willis, S. Rouhani, M. Steinhoff, and M. Baker. 2015. Impacts of the Deepwater Horizon Oil Spill on the salt marsh vegetation of Louisiana. DWH NRDA Shoreline Technical Working Group Report. Prepared for National Oceanic and Atmospheric Administration
- Jacquez, G.M., Goovaerts, P., Kaufmann, A. and R. Rommel. 2014. *SpaceStat 4.0 User Manual: Software for the Space-Time Analysis of Dynamic Complex Systems*, 04/2014; Edition: Fourth Edition, Publisher: BioMedware.
- Løland, A., and G. Høst. 2003, Spatial covariance modelling in a complex coastal domain by multidimensional scaling. *Environmetrics*, **14**(3): 307–321.
- Nixon, Z. 2015. Deepwater Horizon shoreline oil exposure mapping and database: DWH NRDA Shoreline Technical Working Group Report. Prepared for National Oceanic and Atmospheric Administration by RPI.
- Pasta, D.J. 2009. Learning when to be discrete: continuous vs. categorical predictors. *Proceedings of the SAS Global Forum 2009*, 248-2009 http://support.sas.com/resources/papers/proceedings09/248-2009.pdf
- Remy N., Boucher A. and J. Wu. 2008. *Applied Geostatistics with SGeMS: A User's Guide*. Cambridge University Press.
- SAS Institute Inc. 2011. *SAS/STAT 9.3 User's guide*. Cary, NC: SAS Institute Inc.
- Silliman, B., Q. He, P. Dixon, C. Wobus, J. Willis and M. Hester. 2015. Accelerated marsh loss following the BP Deepwater Horizon oil spill: a region wide survey. DWH NRDA Shoreline Technical Working Group Report. Prepared for the Louisiana Coastal Protection and Restoration Authority.
- Swets, J.A. 1988. Measuring the accuracy of diagnostic systems. *Science*, **240**: 1285-1293.

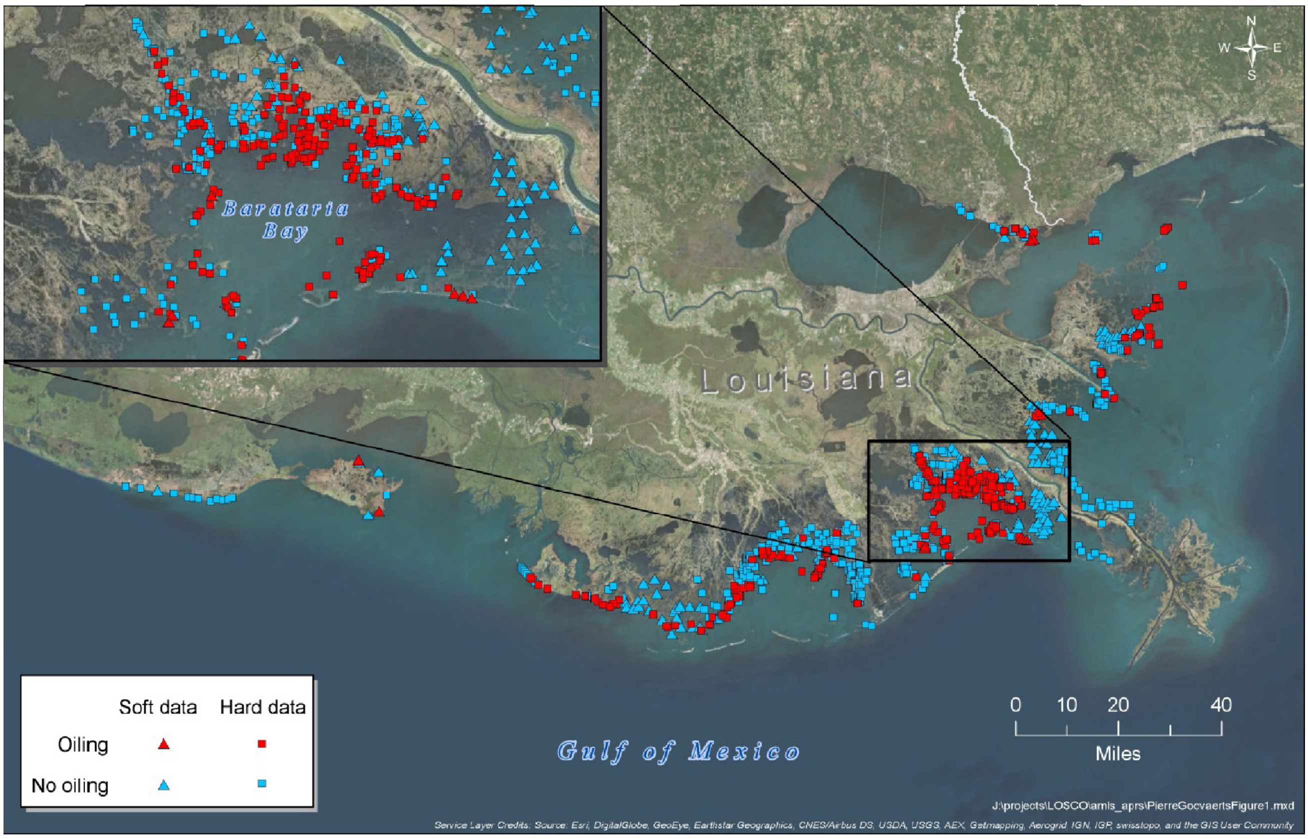
Barataria Bay
N
W
E
S
Louisiana
Soft data
Hard data
Oiling
No oiling
Gulf of Mexico
0
10
20
40
Miles
J:\projects\LOSCO\amls_aprs\PierreGocvaertsFigure1.mxd
Service Layer Credits: Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 1. Geographical location of hard (■) and soft (▲) data on percentage of plant stem oiling that were used in the geospatial analysis. Red color denotes locations where oiling was observed while blue symbols correspond to zero percentage of plant stem oiling.

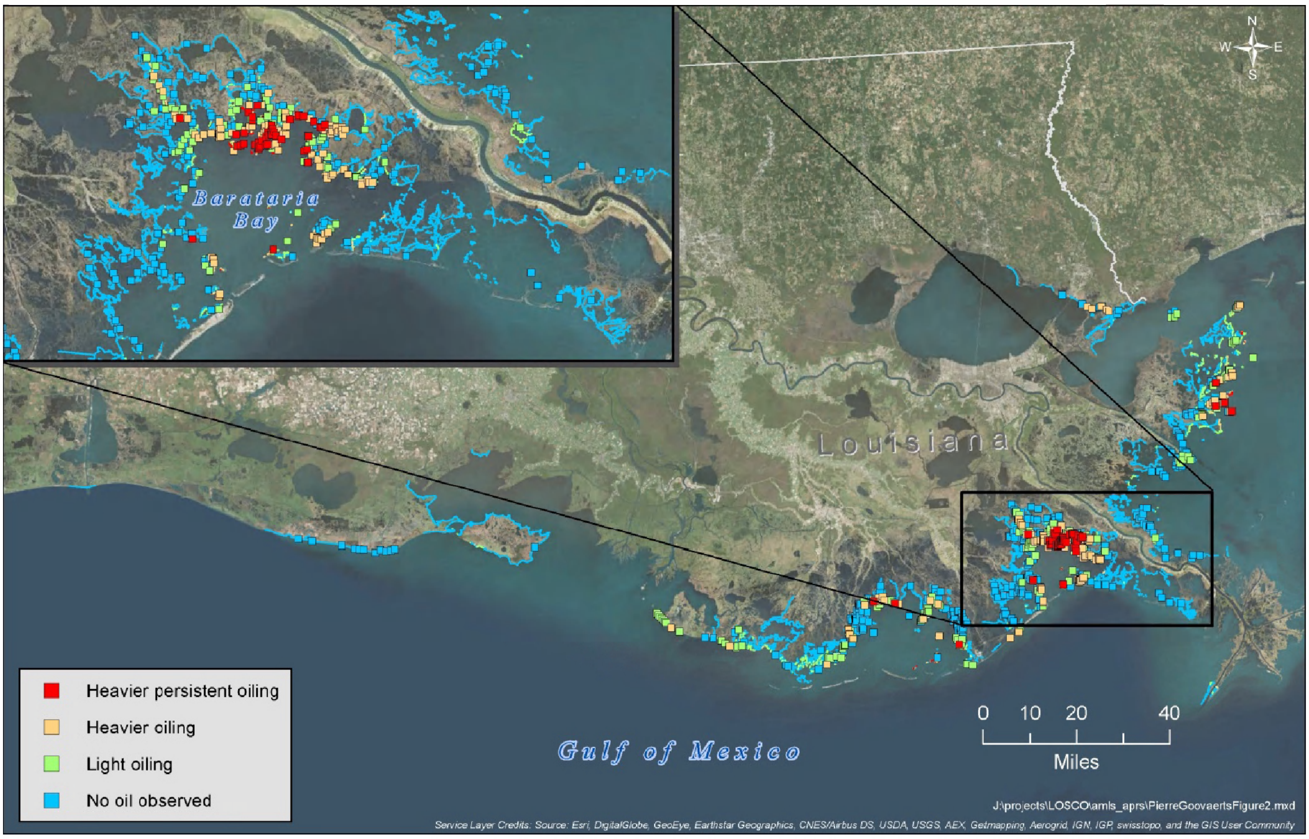
Barataria Bay
N
W
E
S
Louisiana
🟥 Heavier persistent oiling
🟧 Heavier oiling
🟩 Light oiling
🟦 No oil observed
Gulf of Mexico
0
10
20
40
Miles
J:\projects\LOSCO\amls_aprs\PierreGoovaertsFigure2.mxd
Service Layer Credits: Source: Esri, DigitalGlobe, GeoEye, Earthstar Geographics, CNES/Airbus DS, USDA, USGS, AEX, Getmapping, Aerogrid, IGN, IGP, swisstopo, and the GIS User Community

Figure 2. Oiling exposure category recorded at 729 pre-assessment sites, as well as at 118,151 nodes discretizing the Louisiana shoreline under mainland herbaceous marsh. Data frame as in Figure 1.

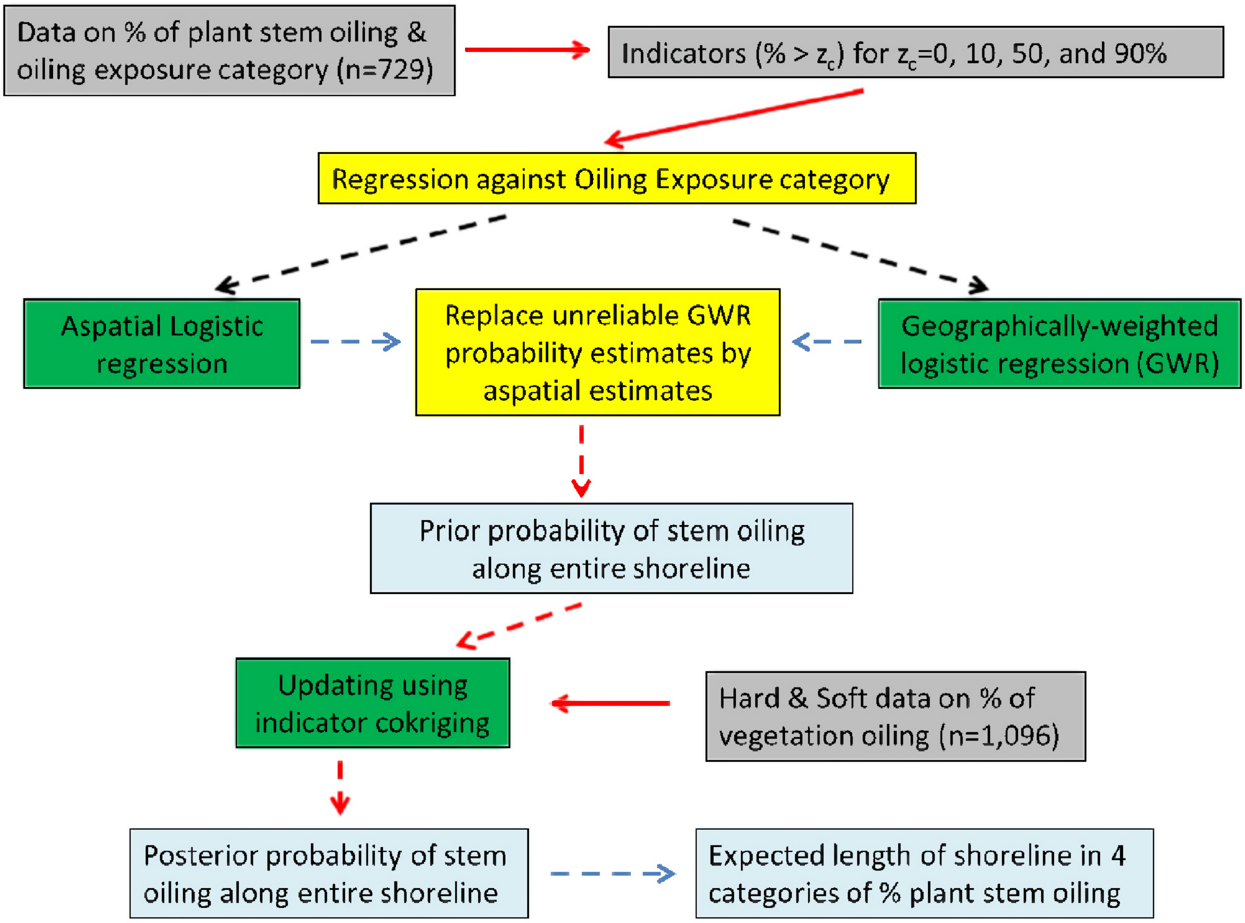
Data on % of plant stem oiling & oiling exposure category (n=729)
Indicators (% > z_c) for z_c=0, 10, 50, and 90%
Regression against Oiling Exposure category
Aspatial Logistic regression
Replace unreliable GWR probability estimates by aspatial estimates
Geographically-weighted logistic regression (GWR)
Prior probability of stem oiling along entire shoreline
Updating using indicator cokriging
Hard & Soft data on % of vegetation oiling (n=1,096)
Posterior probability of stem oiling along entire shoreline
Expected length of shoreline in 4 categories of % plant stem oiling

Figure 3. Flowchart describing the different steps of the geospatial analysis for the computation of expected length of shoreline falling into specific classes of percentage of plant stem oiling.

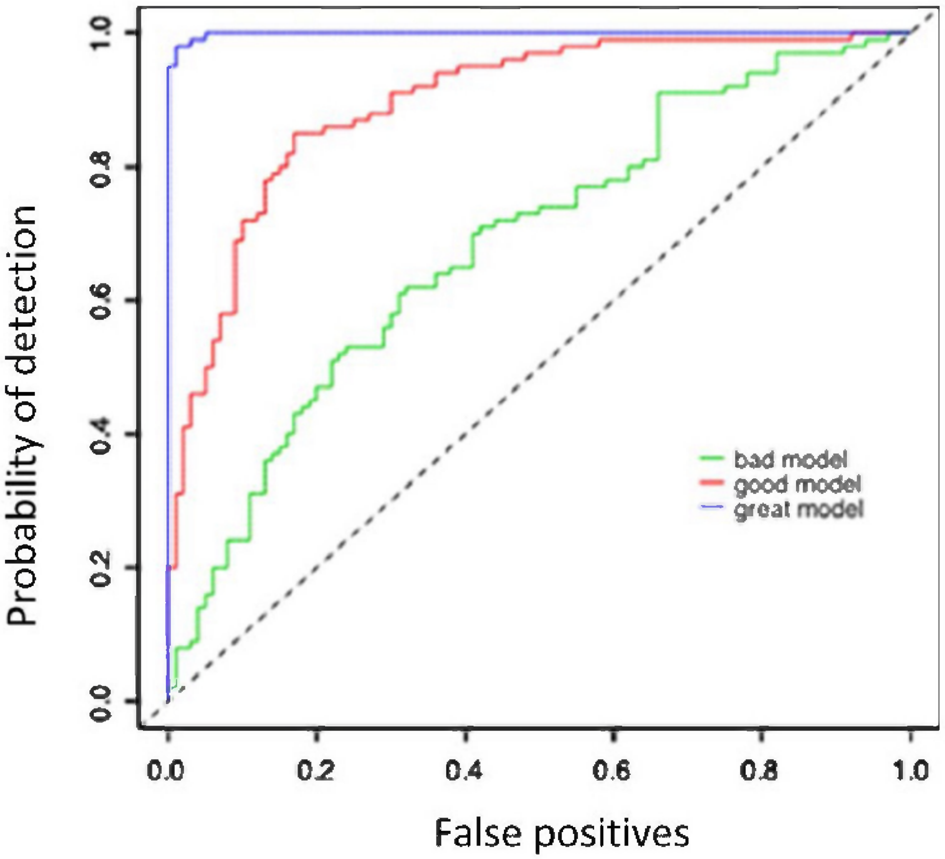

Figure 4. Examples of ROC curves which plot the probability of false positive versus the probability of detection. The most efficient algorithm is the one that allows the detection of a larger fraction of oiled sites at the expense of fewer false positives; that is the ROC curve should be as close as possible to the vertical axis (i.e. blue model). A quantitative measure of the classification accuracy is the relative area under the ROC curve (AUC), which represents the average frequency of detection.

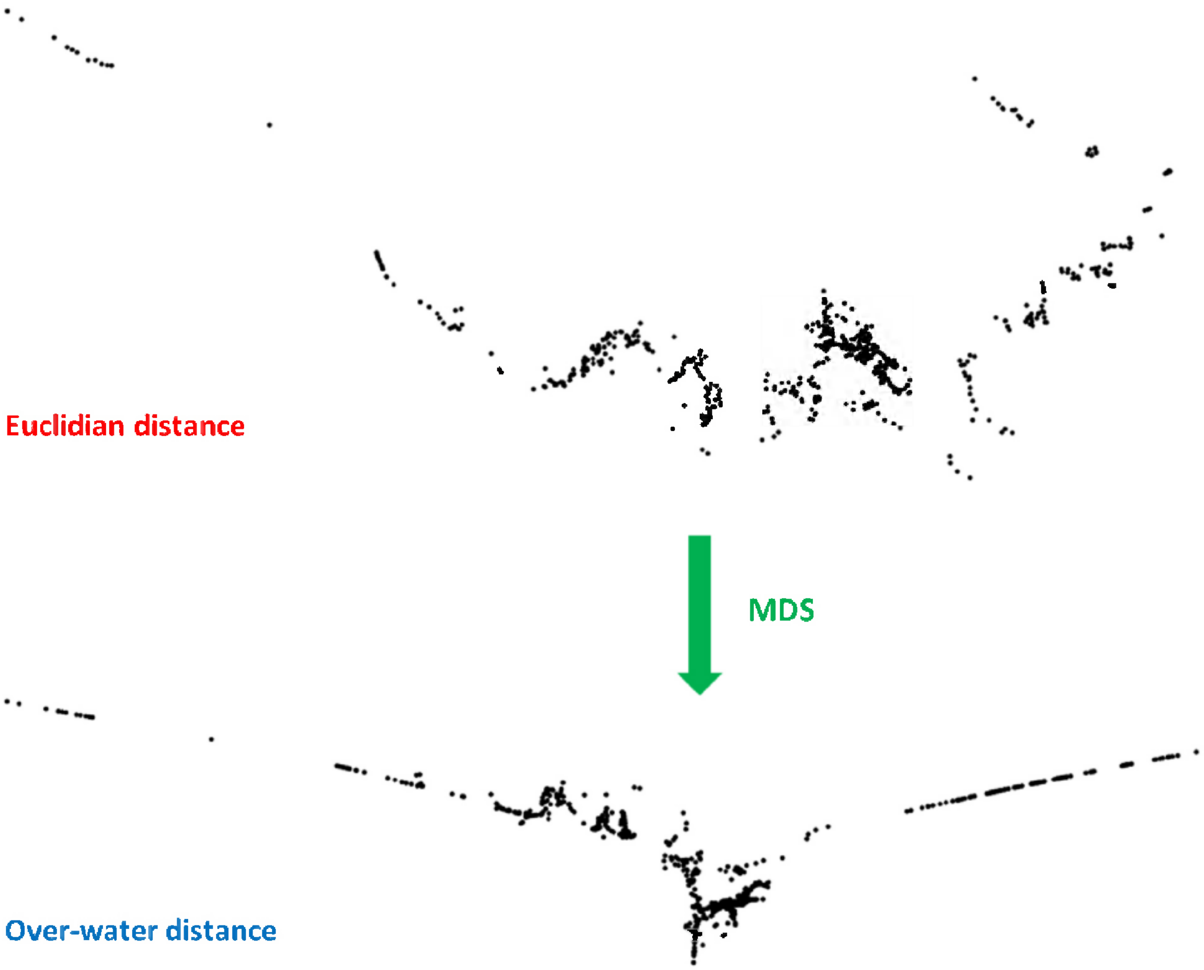

Figure 5. Geometric projection of the original set of 729 pre-assessment locations with known oiling exposure category to create a new data configuration (bottom map) where the Euclidean distance between observations approximates the over-water distance in the original configuration (top map). The approach is based on multidimensional scaling (MDS) and the very small badness-of-fit criterion (0.01) indicates a very good reproduction of water-path distance by Euclidian distance after projection.