Conference on Applied Statistics in Agriculture 2004 - 16th Annual Conference Proceedings

A COMPARISON OF SPATIAL PREDICTION METHODS USING INTENSE SPATIALLY-ACQUIRED WATER QUALITY DATA

E. Barry Moser

Victor H. Rivera-Monroy

Ariel R. Alcantara-Eguren

A COMPARISON OF SPATIAL PREDICTION METHODS USING INTENSE SPATIALLY-ACQUIRED WATER QUALITY DATA

E. Barry Moser, Department of Experimental Statistics, Louisiana State University AgCenter, Baton Rouge, LA 70803-5606

Victor H. Rivera-Monroy, Wetland Biogeochemistry Institute and Department of Oceanography and Coastal Sciences, Louisiana State University, Baton Rouge, LA 70803-5606

Ariel R. Alcantara-Eguren, Departamento de Ciencia e Ingenierias, Universidad Iberoamericana-Puebla, Mexico

Abstract

Water quality information obtained through intensive spatial sampling using automated devices provides opportunities to monitor and forecast the spatial distribution of nutrients and phytoplankton concentrations, and help establish water circulation patterns in estuarine and coastal waters. To be cost effective, efficient sampling designs and estimation methodologies must first be developed. As a starting basis, we applied an original transect sampling design that was used to estimate the spatial distribution of chlorophyll a, salinity, and temperature in the Cienaga Grande de Santa Marta, a coastal lagoon in Colombia. We superimposed the transects over satellite images of the lagoon obtained in the period 1993-2001 to evaluate the efficiency and accuracy of using such transects to estimate the distribution of water quality variables. The satellite images were taken in 1993 (SPOT-3), 1995 (Landsat-6), and 1999 (Landsat-6), and water reflectance values were used as a “proxy” for the water quality variables. Spatial prediction using kriging and thin-plate smoothing splines were used to predict reflectance for a grid network of points taken from the images, and predictions were compared with observed values to compare methods and transect routes. Rapid changes in reflectance in short distances (for example , caused by phytoplankton blooms), complicated the analysis, and neither method proved superior over all transect routes and images, although the kriging predictor remained relatively consistent in performance over the various selected sampling routes.

1. Introduction

A water quality sampling strategy was employed in the Cienaga Grande de Santa Marta estuary system in Colombia, South America, to measure chlorophyll a, salinity, and temperature in the system at monthly intervals from March 19, 1999 through February 28, 2001. The system itself had been impacted through levees and water diversion structures and interest is in reestablishing the system to the extents possible to its prior state (Twilley et al 1998). To help understand how the system has been impacted, and how changes to water management might mitigate the impacts mentioned, knowledge of current salinity, temperature, and primary productivity levels

of the system are required. As the system is dynamic, measurements through time including seasonal changes, replicated over years is also required. Rueda (2001) previously studied this estuary system using systematic sampling with 115 sampling locations spaced 2 km apart on a grid network sampled once during each major season.

To permit large-scale water sampling within a short interval of time, thus providing a "snap-shot" of the system, a high-speed mapping and water sampling device was used. This sampling device consisted of a small boat outfitted with a geographic positioning system (GPS) device, a water pass-through system with sensors that conducts fluorometry (used to measure chlorophyll a) and salinity measurements at 10 second intervals, and a data logger to record geographic position and the water quality measurements. In this particular system, water temperature was recorded manually at 1 minute intervals. By driving the boat along a predetermined route at constant speed, water quality data are collected at very short intervals of space along the routes. Example sampling routes are shown in Figures 1a and 1b.

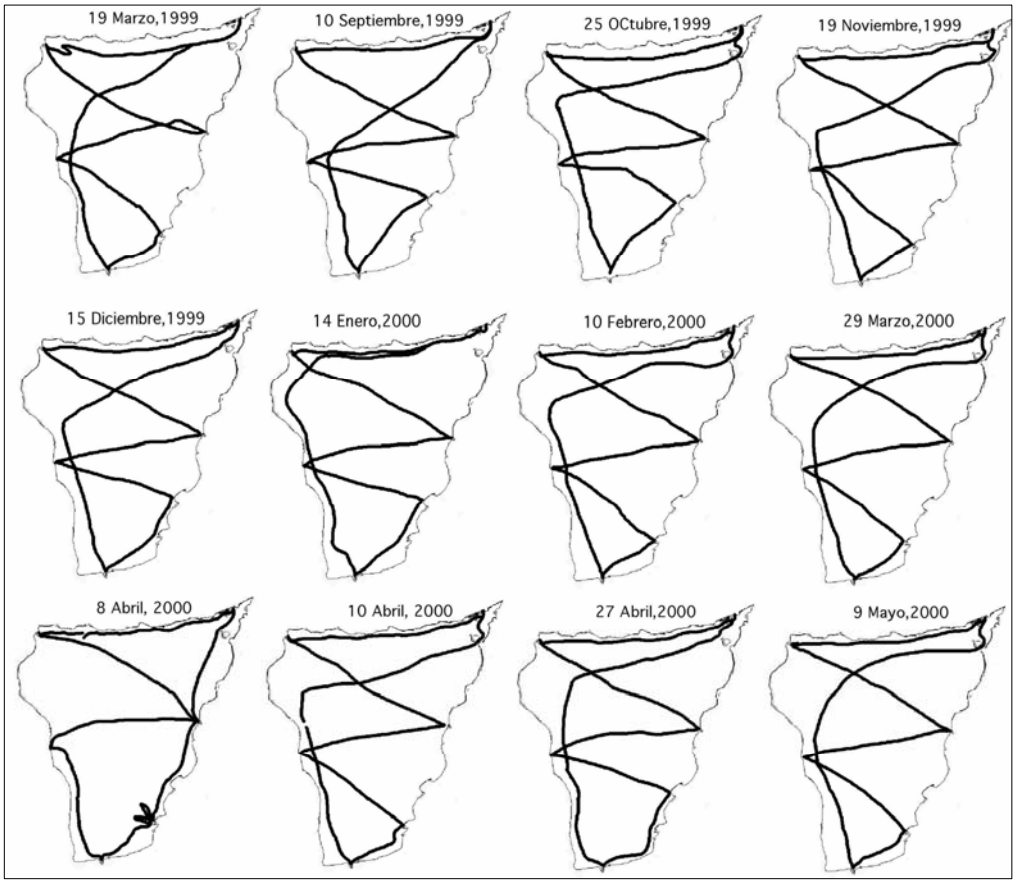

Figure 1a. Sampling routes employed in sampling water quality data from the Cienaga Grande de Santa Marta estuary system (March 19 1999-May 9, 2000).

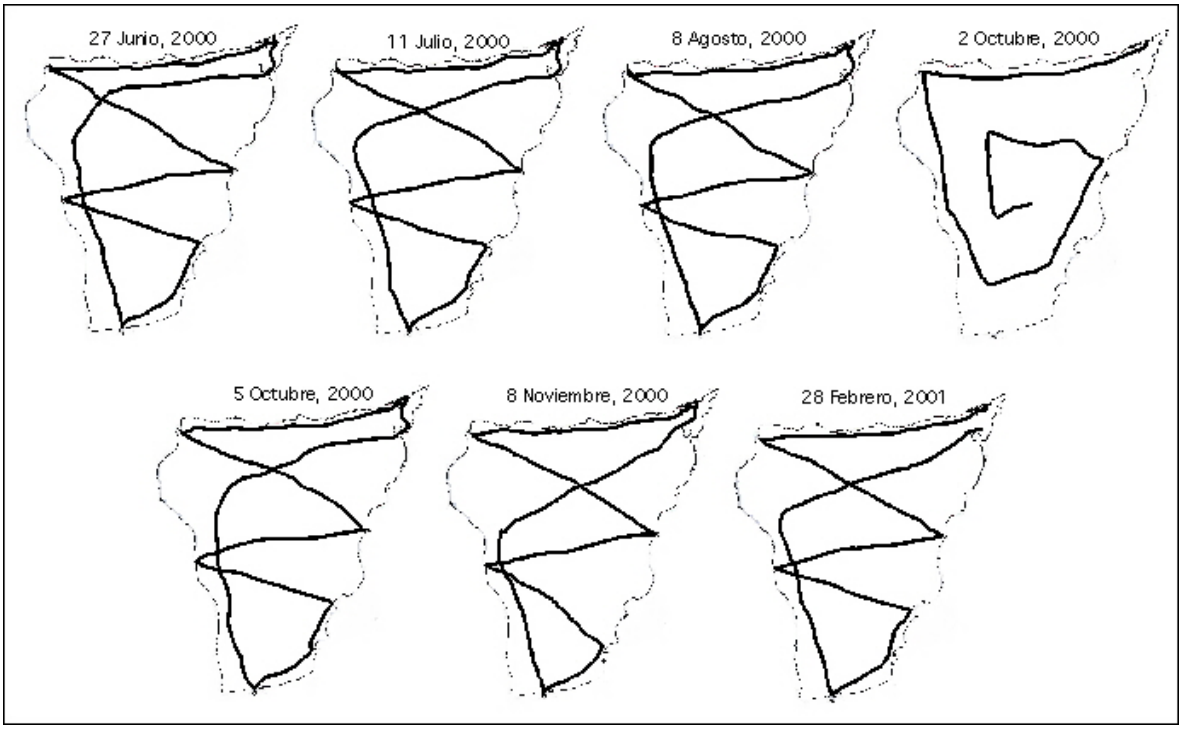

Figure 1b. Sampling routes employed in sampling water quality data from the Cienaga Grande de Santa Marta estuary system (June 27 2000-February 28, 2001).

The challenge for the data analysis is to use data collected along the routes at each of the sampling dates to predict the water quality information for the entire system. In addition, very large or small “outlier” or “spike” values would be observed when the boat crossed particular currents, eddies, and juxtaposed water bodies along the sampling routes. These currents and water bodies are associated with freshwater inputs from rivers emptying into the system, and from saline inputs from the Caribbean Sea.

Spatial data such as the water quality data collected in this study have been analyzed and spatially interpolated or predicted using a variety of methods. In particular, trend surface models using polynomial regressions, ordinary kriging and its generalizations, and smoothing splines have frequently been used (see e.g., Cressie 1991 for a variety of examples). To understand the usefulness of ordinary kriging and of smoothing splines for prediction and interpolation of the water quality data, a simulation study was developed to compare the methods. LANDSAT and SPOT images of the Cienaga Grande de Santa Marta system were available for 1993, 1995, and 1999. These images show spatial variability within the system thought representative of variability that would likely be observed in chlorophyll a, salinity, and temperature measurements. Indeed, data “spikes” are also contained within the images. Our objective was to provide a simple approach with a set of sequential steps to make this analysis as simple as possible (“user oriented”) once the main algorithms were selected. Thus we did not consider specific analysis for the spikes or outliers when their frequency was low. Although in cases when

this frequency of outliers/spikes is high further analyses (post-stratification, de-trending) might need to be included.

2. Simulation Methods

Reflectance values were sampled from each geo-referenced SPOT 1993, LANDSAT 1995, and LANDSAT 1999 image. Although the images do not represent true replicates since each image was captured during large scale climatic events (El Niño Southern Oscillation/La Niña), each image still represents a point of reference to assess changes in reflectance at a regional scale for each individual date. This inter-annual variability is part of the system behavior and needs to be incorporated in the analysis. Each sampling route (19 routes) was superimposed upon each image and the reflectance for the nearest pixel to each sampling point along each route was retrieved. Thus, the satellite image was treated as "ground truth" data, and the superimposed routes were used to provide simulated samples representing the original water quality samples. An example of such a sampled route is given in Figure 2.

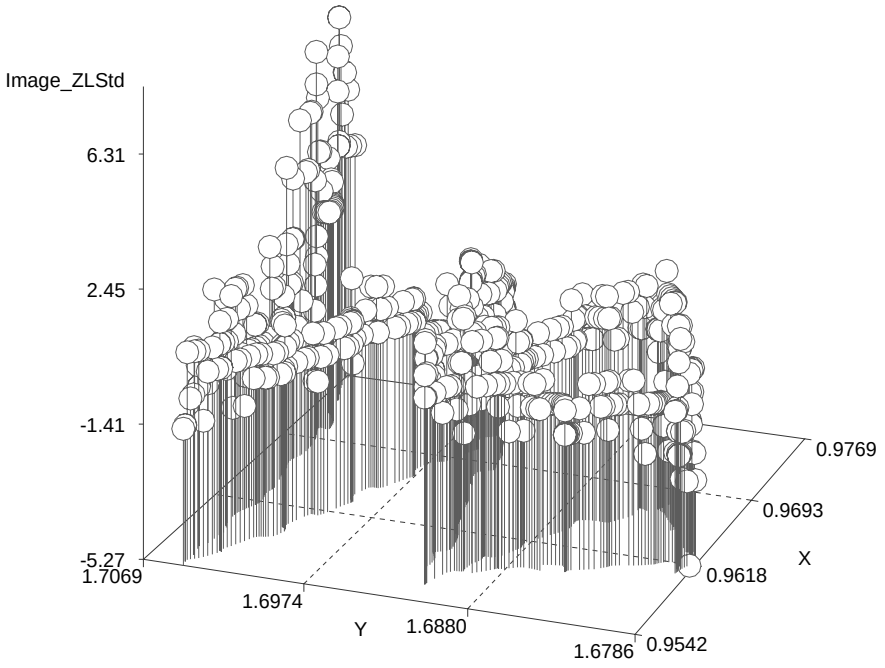

Figure 2. Sampled log standardized reflectance values using the route of November 19, 1999.

A square grid of 2500 points was also placed over each image and the reflectance value at each grid point was obtained. These points would serve as target points for prediction and evaluation of the estimation methods. Any grid points that fell outside of the Cienaga Grande system were excluded. Predictions were made to each grid point and mean square prediction errors were computed for comparisons among routes and between analysis methodologies for each image. Data were analyzed on the observed and log-transformed scales, and with and without the outlier

or “spike” data included. Outliers were detected as points outside of the outer fence of a box-plot of all of the reflectance values in the sampled route.

PROC VARIOGRAM of the SAS System was used to estimate empirical robust variograms for the sampled track data. Exponential and spherical variogram models were then fit to the empirical robust variograms using PROC NLIN, though we have reported only the results using the exponential variogram herein, as there were not great differences between kriging predictions between the variogram types. An example fitted variogram is given in Figure 3.

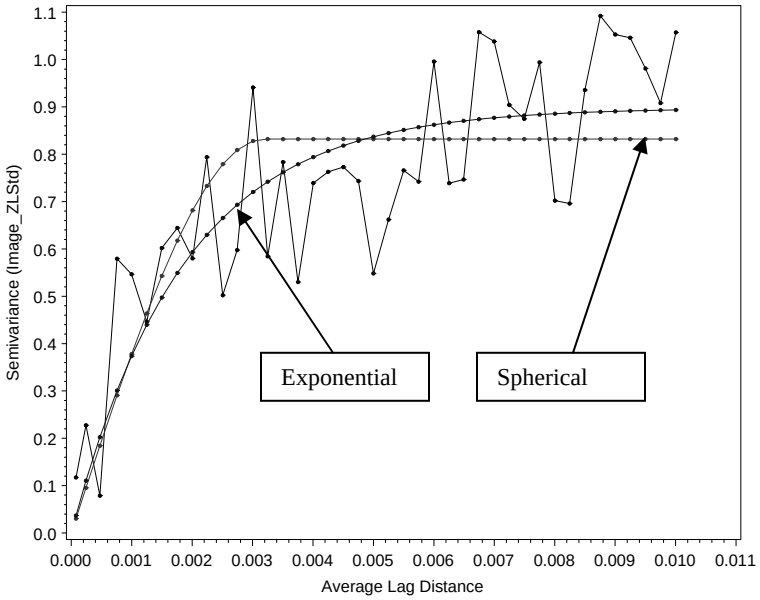

Figure 3. Fitted exponential and spherical variogram models superimposed upon the robust empirical variogram. Only the exponential variogram was used in this simulation study.

Resulting parameter estimates of the variogram models were then passed to PROC MIXED for kriging prediction to the grid points assuming a constant mean structure. PROC TPSLINE was used to fit thin-plate smoothing splines to the surface generated from each of the sampling routes. Default procedure settings were used to control the fit of the spline models to the data. By default TPSLINE selects its smoothing parameter using generalized cross validation, and the order of the derivative for the penalty function for the penalized least-squares estimate is max(2, INT(d/2)+1), where d is the number of smoothing variables (SAS Institute Inc. 2005).

In addition, characteristics of each sampling route or track were computed and included the number of sample points in the track, the total length of the track (sum of all track segment lengths), and the largest nearest neighbor distance from grid points to track points. It was thought that longer tracks and tracks that minimized the largest nearest neighbor distance would tend to have superior prediction properties, and therefore, would provide insight into optimal sampling designs with respect to track layouts.

3. Results

Predictions based upon the log transformed data appeared much superior to those using the untransformed data when rescaled. Results herein will focus upon the predictions using the log transformed data. In general, predictions from either method that did not exclude outliers from the training data sets (Table 1) had considerably larger prediction errors than when the outliers were excluded (Table 2). The predictions produced using the thin-plate splines tended to produce a more complex surface than when using kriging (Figure 4).

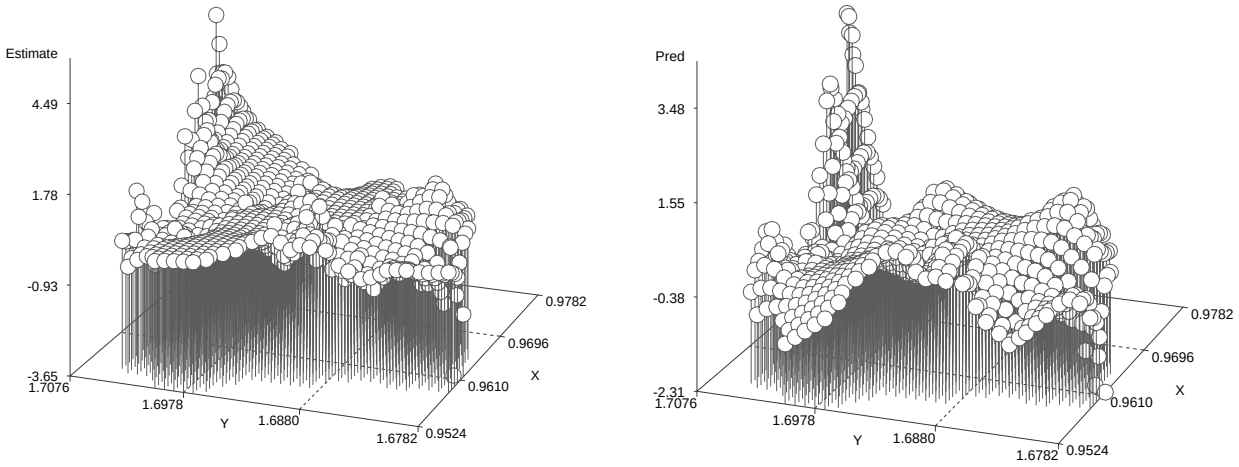

Figure 4. Kriging (left) and thin-plate smoothing spline (right) predictions of reflectance for the LANDSAT 1999 image using the November 19, 1999 sampling route.

The kriging estimator produced more consistent results (similar mean-squared errors of prediction) across the different sampling tracks when sampling from the same image, than did the thin-plate smoothing spline predictor. For the 1993 and 1995 images with outliers included, the sampling route of 02OCT2000 produced a very large prediction error relative to other routes for the thin-plate smoothing spline, while the kriging estimator showed little variation in prediction performance for this route when compared to the other routes. Note that this route is quite different in its spatial arrangement relative to the other routes (Figure 1b).

The mean-squared errors of prediction from the various routes, images, and prediction methods were then related to the route characteristics. No associations were noted between the mean-squared prediction errors and route characteristics. It was thought that the maximum nearest neighbor distance of the grid points to the track might be most important though the observed relationship was noisy (Figure 5).

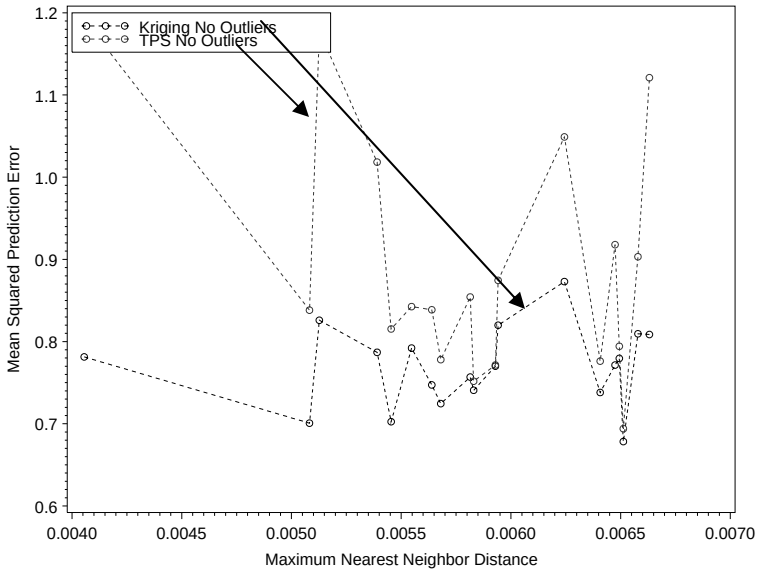

Figure 5. Mean-squared prediction error of kriging and thin-plate smoothing spline predictors for the 1999 LANDSAT image using all routes as a function of the maximum nearest neighbor distance of the grid points to the sampled route points. Outliers were excluded from the route data prior to the computation of the predicted values.

4. Discussion

The water quality data collected during the actual study contains outliers or “spikes” at irregularly-spaced intervals along the sampled routes. The same phenomenon was observed when sampling from the SPOT and LANDSAT images. The results from this simulation study suggest that removing these outliers prior to surface prediction is important in order that the major surface trends are correctly predicted. Further the thin-plate smoothing spline predictions appeared to be more sensitive to the outliers than those of the kriging estimator. Alternative models will need to be explored if prediction of the “spike” process is desired. Overall the kriging estimator outperformed the thin-plate smoothing spline estimator for surface prediction for the routes and images considered in this study. Ramsay (2002) proposes a penalized smoothing bivariate finite-element process to deal with interior holes and irregular boundaries and found that his technique worked better than thin-plate splines for two examples that he considered. These sampling routes do leave holes in the domain and the boundaries are somewhat irregular. Thus, the finite-element technique of Ramsay (2002) should be investigated for these data. O’Connell and Wolfinger (1997) found little difference between the performance of kriging and thin-plate spline predictors in their simulation study, though both of these methods outperformed polynomial trend-surface models. Laslett (1994) contends that kriging never performs worse than splines, but that sometimes it greatly outperforms spline prediction. He finds that the sampling regime determines when they are similar or when kriging will work much better, and states that clustered sampling sites favor the kriging estimator over splines. Our study supports the findings of Laslett (1994) and strongly suggests that analyses and future sampling routes be based upon kriging estimation. Local kriging and detrending methods, including post-

stratification, warrant study as well as we are predicting over a large geographic area relative to the spatial scale of sampling.

Track or route characteristics such as length of track, number of points in the sample, and maximum nearest neighbor distance of the grid points to the sampled route points appear not to be associated with prediction performance. However, the track layouts are all quite similar and so alternative tracks not employed in the actual survey should be added into this simulation study. It appears that the spline estimator is much more sensitive to the route taken than is the kriging estimator, though the characteristics measured do not indicate the exact nature of this relationship.

5. Summary

The use of sampled satellite imagery to mimic intense spatially-acquired water quality data appears to generate data that have characteristics much like observed water quality data, including outliers and rapidly changing contours. This known imagery data is then used to help validate approaches to data analysis that practitioners of precision-sampled water quality data are likely to use. Our findings suggest that an ordinary kriging approach using a robust variogram estimate is more likely to result in smaller mean square prediction errors than would a spline fitted surface to the data. Additional research into optimal sampling trajectories with respect to these estimation methods, and exploration of other spatial modeling methodologies is needed to help identify the most robust approaches for this type of data.

6. References

Cressie, N. 1991. *Statistics for Spatial Data.* John Wiley & Sons, Inc., New York, N.Y., 900pp.

Laslett, G. M. 1994. Kriging and splines: an empirical comparison of their predictive performance in some applications. *J. Amer. Statist. Assoc.* 89 (426), 391-400.

O'Connell, M. A. and R. D. Wolfinger. 1997. Spatial regression models, response surfaces, and process optimization. *J. Comp. and Graph. Statistics* 6 (2), 224-241.

Ramsay, T. 2002. Spline smoothing over difficult regions. *J. R. Statist. Soc. B* 64, part 2, 307-319.

Rueda, M. 2001. Spatial distribution of fish species in a tropical estuarine lagoon: a geostatistical appraisal. *Mar. Ecol. Prog. Ser.* 222, 217-226.

SAS Institute, Inc. 2004. *SAS OnlineDoc 9.1.3.* SAS Institue, Cary, NC. http://support.sas.com/onlinedoc/913/docMainpage.jsp.

Twilley, R. R., Rivera-Monroy, V. H., Chen, R., Botero, L. 1998. Adapting an ecological mangrove model to simulate trajectories in restoration ecology. *Marine Pollution Bulletin* 37: 404-419.

Table 1. Mean-squared prediction error (MSPE) for kriging and thin-plate smoothing spline (TPS) predictors for the 1993, 1995, and 1999 satellite images for the 19 sampling routes using log transformed reflectance and including outliers in the analysis.

Sampling Route Date	Kriging MSPE 1993	TPS MSPE 1993	Kriging MSPE 1995	TPS MSPE 1995	Kriging MSPE 1999	TPS MSPE 1999
19-Mar-99	0.962	1.242	1.820	1.815	0.819	0.830
10-Sep-99	1.558	2.459	2.212	2.424	0.939	1.418
25-Oct-99	2.003	1.855	3.021	3.183	0.962	1.562
19-Nov-99	2.423	**5.642**	2.575	2.902	0.954	2.511
15-Dec-99	1.675	1.710	2.622	2.527	0.807	1.078
14-Jan-00	1.923	**7.721**	2.670	2.538	0.835	3.138
10-Feb-00	1.319	1.962	2.722	3.218	1.010	2.618
29-Mar-00	3.012	2.762	2.714	2.888	1.301	1.383
8-Apr-00	3.187	3.431	3.009	3.485	1.282	1.624
10-Apr-00	2.590	2.784	2.252	2.216	1.207	1.499
27-Apr-00	1.447	**11.316**	2.372	2.672	0.933	**5.170**
9-May-00	3.518	3.413	2.780	2.821	1.426	1.705
27-Jun-00	3.005	2.951	2.373	2.555	1.275	1.379
11-Jul-00	3.408	3.279	2.727	2.566	1.310	1.437
8-Aug-00	2.665	3.841	2.664	2.620	1.234	1.852
2-Oct-00	2.749	**21.290**	2.743	**4.288**	0.915	**5.167**
5-Oct-00	2.486	2.416	2.653	2.845	1.259	1.379
8-Nov-00	2.313	2.790	2.445	2.437	1.316	1.984
28-Feb-01	2.237	2.460	2.636	2.570	1.123	1.342

Table 2. Mean-squared prediction error (MSPE) for kriging and thin-plate smoothing spline (TPS) predictors for the 1993, 1995, and 1999 satellite images for the 19 sampling routes using log transformed reflectance but excluding outliers prior to the analysis.

Sampling Route Date	Kriging MSPE 1993	TPS MSPE 1993	Kriging MSPE 1995	TPS MSPE 1995	Kriging MSPE 1999	TPS MSPE 1999
19-Mar-99	1.010	1.068	2.238	2.248	0.792	0.843
10-Sep-99	0.961	0.951	2.269	2.477	0.678	0.694
25-Oct-99	1.009	0.954	2.528	2.612	0.781	1.192
19-Nov-99	0.969	1.142	2.575	2.902	0.809	1.121
15-Dec-99	0.848	0.800	2.594	2.577	0.701	0.838
14-Jan-00	0.865	1.012	2.528	2.433	0.703	0.815
10-Feb-00	0.905	0.990	2.557	2.661	0.747	0.839
29-Mar-00	0.947	0.937	2.420	2.558	0.780	0.794
8-Apr-00	1.083	1.454	2.859	3.192	0.771	0.918
10-Apr-00	0.869	0.913	2.332	2.331	0.724	0.778
27-Apr-00	0.938	1.436	2.416	2.752	0.826	1.178
9-May-00	0.953	1.043	2.634	2.667	0.873	1.049
27-Jun-00	0.986	1.094	2.433	2.553	0.770	0.772
11-Jul-00	1.190	1.040	2.360	2.438	0.738	0.776
8-Aug-00	0.924	0.966	2.530	2.542	0.757	0.854
2-Oct-00	1.058	1.326	2.650	**4.011**	0.820	0.875
5-Oct-00	0.850	0.895	2.525	2.690	0.809	0.903
8-Nov-00	0.862	0.851	2.379	2.443	0.787	1.018
28-Feb-01	0.945	0.935	2.536	2.507	0.741	0.752