CS229 Project: Identifying Regions of High Turbidity in San Francisco Bay

Joe Adelson

December 11, 2014

Introduction

Suspended sediments in oceans, seas, and estuaries shape coastal geography, provide important nutrients to ecosystems, and transport and bury harmful contaminants. Although the problem has significant interest in the scientific and environmental engineering communities, many of the the mechanisms involved are poorly understood because of the difficulty and expense of measuring what turns out to be a very complex system. The complexity arises from the fact that many estuaries are dynamic regions, where the currents that move sediment depend on rainfall, wind, waves, tides, salinity gradients, and anthropogenic manipulation (dams, weirs, etc.). On top of this the physics of how sediment particles interact with the sea bed and one another is still poorly understood, especially at the scales of interest to biologists and coastal engineers. Because of these challenges the development of tools to study the problem is an active area of research.

The traditional tools of studying sediment transport are in situ field measurements and numerical models, and these each face challenges. Because of the small scale of in situ sampling methods, it is often difficult (and very expensive) to collect data over a large enough area to develop a strong regional understanding of what is happening. Numerical models have thus gained popularity as a way to predict suspended sediment concentrations (SSC), erosion, and deposition, but these contend with their own problems of unknown boundary conditions and poorly understood transport physics for silts and clays. To help fill the gap left in these two technologies, there has been work to develop remote sensing algorithms using available NASA and European Space Agency (ESA) satellite imagery to map turbidity, an optical measure of the cloudiness of water, which serves as a proxy for SSC. This is still a young technology and has substantial room for development.

San Francisco Bay is an excellent example of a system that showcases many of importances of sediments and the challenges with understanding them. For one, it is a well studied estuary with ample of measurements about the flow and sediment conditions publicly available.

This project's goal is to establish a relationship between available remote sensing data and in situ measurements of turbidity in San Francisco Bay. This requires correlating the data at the particular pixel of the measurement (and perhaps its neighbors) with the measurement itself. This has been studied in other estuaries [1] and there are products available that can calculate turbidity, but these are not calibrated for our area of interest.

Data Collection

There are two primary sources of data for this project: remote sensing satellite data and in situ point measurements. Satellite images from the MERIS probe taken in the time period of 2006 to 2012 were downloaded from the CoastColour project [3], a data offering from the ESA that specializes in processing coastal images. Each image pixel contains information of the intensity of discretized reflectance of both visual and infrared wavelengths, as well as precomputed estimates of turbidity, suspended matter, pigment, and chlorophyll. Because the turbidity is of significant

interest, a spatially averaged turbidity is calculated in an attempt to create a turbidity feature with reduced noise.

In situ measurements of turbidity are taken from three United States Geological Survey (USGS) monitoring stations at Alcatraz Island, the Dumbarton Bridge, and the Richmond-San Rafael Bridge each taken at a depth of 6 to 8 meters below the surface. These measurements are taken every 15 minutes.

Preprocessing the data includes extracting useful information for both datasets at matching times and locations. This includes removing cloudy images as well as finding the image pixel that contains the USGS sampling station. In all there are at most three samples for each image, one for each valid image pixel and station measurement combination, for a total of 679 samples of 44 features each.

Regression Methods

Linear regression with L1 and L2 shrinkage parameters as well as support vector regression were tested. The examined parameters are: number of principal components, degree of polynomial expansions of the feature space, and SVR kernels. Many tests were run on the dataset to find the optimal values of these numbers. For each test a parameter sweep for the optimal penalty logarithmic parameter was done using K-fold cross validation with 5 folds.

In order to ensure that the turbidity levels are non-negative, all regressions are completed using the log of the turbidity values. The performance metrics are given using the actual turbidity units

The computations use the Python library Scikit-Learn library [2]. The limiting factors for testing polynomial expansions and number of principal components were both the run time of the parameter sweeps and the trend towards overfitting with high polynomial degrees and many features. Therefore, polynomial regression with high degrees was limited to using relatively few principal components.

Results

The linear regressions perform better than support vector regression (table 1). The high order polynomials are generally optimal with very large shrinkage parameters, which implies that high order polynomials overfit the data. Below are mean RME associated with the optimal penalty parameter for some of the tested PCA and polynomial combinations tested via 5-folds cross validation (figures 1, 2, 3). As a point of comparison the CoastColour turbidity measure has a root mean square error (RME) of 108.5. Although this work shows substantial improvement CoastColour sediment estimate, the R-Squared fit is only 0.22.

PCA does not appear to be an effective tool for eliminating overfitting of the data as the full set of data performed best. Also, polynomial expansion of the features did not improved RME. The regressions are sensitive to the shrinkage parameter and the search finds a smooth minimum for the linear regressions (figure 4) but SVR is more erratic (figure 5). These regressions tend to under predict the large measured turbidities (figure 6).

Discussion

We are able to make a significant improvement over the "out-of-the-box" measure of turbidity for San Francisco Bay. However, the optimal test set R-Squared value of 0.22 suggests that there is much work to be done in predicting the turbidity. The reason for the poor performance likely occurs for a variety of reasons. Most apparent is the disparity of scale in the data we use: Because the satellite data has a spatial resolution of about 300 meters, it will inevitably not be able to pick up the small scale feature that may affect the in situ measurements. Secondly, there is a three dimensionality to turbidity and the satellite reads the surfaces, while the in situ measurements are sampled at a depth of 6-8 meters depending on the tide. There may also be non-linearities between the satellite data and the actual turbidity that our regression models do not pick up. Future work includes not only expanding this dataset to more features and testing the value of neural networks, but also a proposal to conduct our own experiment of taking aerial photographs of the bay, while measuring the turbidity levels of South San Francisco Bay using boat transects to get a wider spatial baseline of measurements.

References

[1] R. L. Miller and B. A. McKee. Using {MODIS} terra 250 m imagery to map concentrations of total suspended matter in coastal waters. *Remote Sensing of Environment*, 93(1):259 – 266, 2004.

[2] F. e. a. Pedregosa. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[3] K. Ruddick, C. Sa, S. Bernard, L. Robertson, M. Matthews, R. Doerffer, W. Schoenfeld, H. Z. G. HZG, M. S. Salama, S. Budhiman, et al. Coastcolour round robin protocol in situ reflectance data set.

Table 1: Best performance for L1, L2, and Support Vector Regressions

Estimate	RME	Optimal Parameter
CoastColour Turbidity	108.5	N/A
L2 Regression	2.20	4.00
L1 Regression	2.20	7.87 $\times 10^{-}4$
Support Vector Regression	2.72	0.054 (RBF Kernel, 5 Principal Components)

Optimal K-Folds (5) RME for Polynomial Regression with L2 Shrinkage							
PCA \ Degree	1	2	3	4	5	6	7
1	2.88	2.88	2.87	2.81	2.71	2.53	2.87
2	2.88	2.87	2.68	2.22	2.88	2.22	2.88
4	2.88	2.85	2.82	2.88	2.87	2.88	
6	2.88	2.81	2.88	2.88	2.87	2.88	
8	2.68	2.25	2.88	2.88			
10	2.67	2.20	2.88	2.88			
15	2.64	2.50	2.88	2.88			
20	2.21	2.56	2.88	2.88			
Full	2.20	2.41	2.88	2.88			

Optimal Parameter for Polynomial Regression with L2 Shrinkage							
PCA \ Degree	1	2	3	4	5	6	7
1	1.2E+21	1.2E+21	6.0E+05	9.8E-04	9.8E-04	1.1E-03	1.2E+21
2	1.2E+21	4.6E+05	9.8E-04	9.8E-04	1.2E+21	1.2E+04	1.2E+21
4	1.2E+21	1.3E+05	5.4E+03	1.2E+21	6.2E+26	6.3E+29	
6	9.8E-04	1.9E+04	1.2E+21	1.2E+21	6.2E+26	6.3E+29	
8	1.6E+04	6.8E+02	1.2E+21	1.2E+21			
10	1.4E+04	3.4E+02	1.2E+21	1.2E+21			
15	9.8E-04	6.0E+05	1.2E+21	1.2E+21			
20	9.8E-04	9.1E+05	1.2E+21	1.2E+21			
Full	4.0E+00	6.0E+05	1.2E+21	1.2E+21			

Figure 1: Optimal shrinkage parameter (via K-Folds optimization) and associated RME for measured and predicted with L2 shrinkage for turbidity (FNU).

Optimal K-Folds (5) RME for Polynomial Regression with L1 Shrinkage							
PCA \ Degree	1	2	3	4	5	6	7
1	2.88	2.88	2.87	2.84	2.83	2.83	2.83
2	2.88	2.87	2.75	2.70	2.70	2.69	2.69
4	2.88	2.86	2.88	2.69	2.81	2.85	
6	2.86	2.83	2.76	2.88	2.85	2.81	
8	2.66	2.27	2.73	2.78			
10	2.66	2.22	2.88	2.84			
15	2.65	2.41	2.48	2.86			
20	2.22	2.40	2.88	2.86			
Full	2.20	2.36	2.86	2.86			

Optimal Parameter for Polynomial Regression with L1 Shrinkage							
PCA \ Degree	1	2	3	4	5	6	7
1	1.3E+00	3.0E+00	1.1E+00	9.8E-04	9.8E-04	9.8E-04	9.8E-04
2	1.3E+00	4.4E-01	9.8E-04	9.8E-04	9.8E-04	9.8E-04	9.8E-04
4	1.3E+00	5.7E-01	2.2E+04	4.4E-01	5.1E+02	2.0E+00	
6	2.5E-01	9.8E-04	3.0E+00	1.8E+03	5.7E-01	5.6E-01	
8	2.2E-01	9.0E-03	1.2E+01	2.6E+02			
10	2.2E-01	2.1E-02	1.9E+04	3.1E+03			
15	9.8E-04	5.0E-01	5.9E-03	3.6E+03			
20	9.8E-04	6.6E-01	1.9E+04	3.6E+03			
Full	7.9E-04	6.3E-02	1.9E+04	1.9E+04			

Figure 2: Optimal shrinkage parameter (via K-Folds optimization) and associated RME for measured and predicted with L1 shrinkage for turbidity (FNU).

Optimal K-Folds (5) RME for Support Vector Regression				
Kernel / PCA	Linear	RBF	Polynomial Degree 2	Polynomial Degree 3
5	2.77	2.72	2.78	2.76
10	2.84	3.13	3.66	3.66
20	3.01	3.63	3.66	3.66
Full	3.62	3.63	3.66	3.66

Optimal Parameter for Support Vector Regression				
Kernel / PCA	Linear	RBF	Polynomial Degree 2	Polynomial Degree 3
5	5.4E-02	1.7E+00	8.2E-02	2.7E-02
10	5.1E+02	8.9E+02	9.8E-04	9.8E-04
20	8.9E+02	6.4E+01	9.8E-04	9.8E-04
Full	5.6E+01	1.3E+02	9.8E-04	9.8E-04

Figure 3: Optimal penalty parameter (via K-Folds optimization) and associated RME for SVR

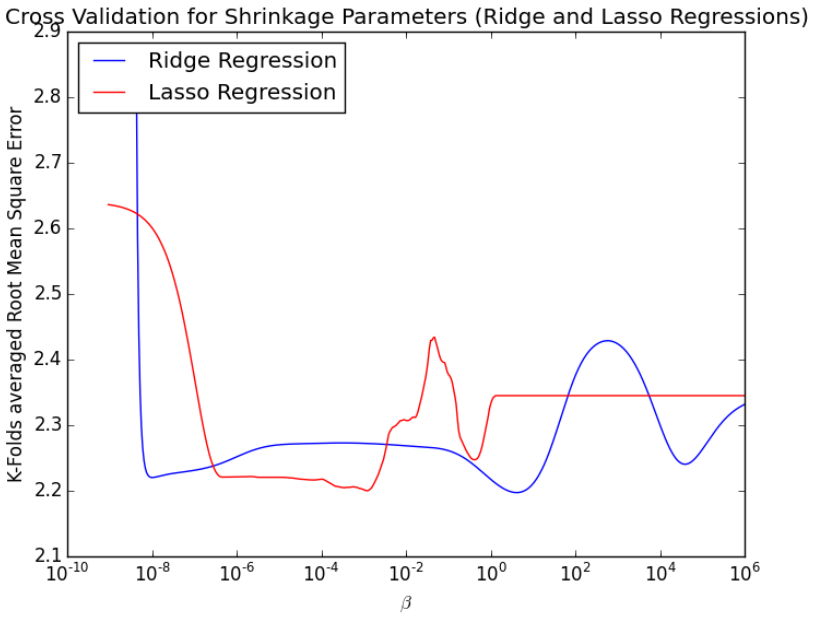

Figure 4: Sweep for the L1 (Lasso) and L2 (Ridge) shrinkage parameters (β).

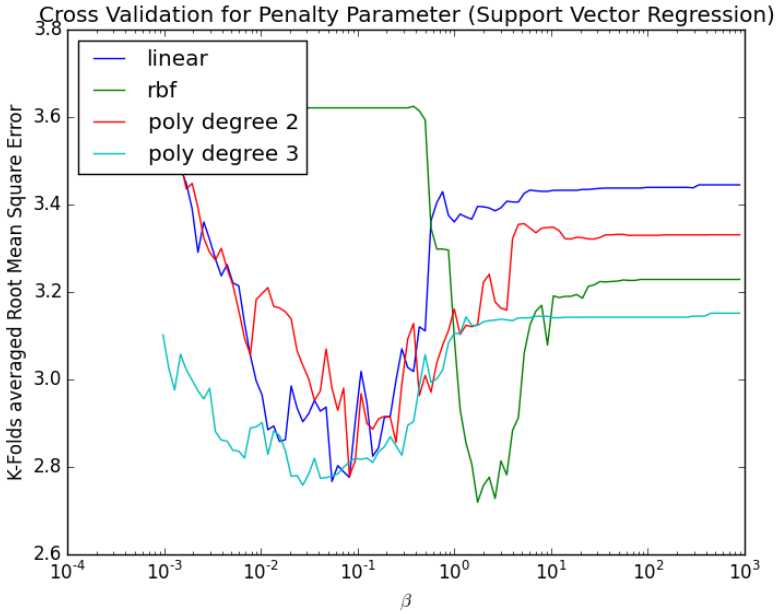

Figure 5: Sweep of the penalty parameter, β for SVR.

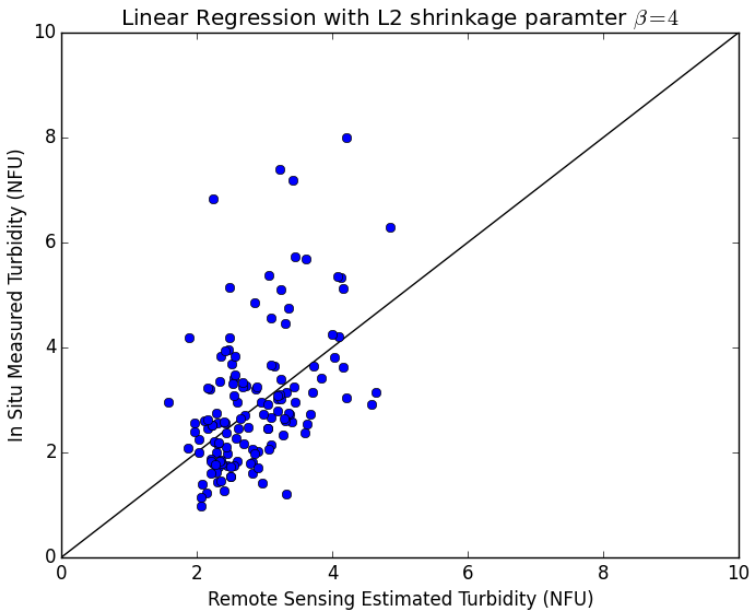

Figure 6: Measured turbidity vs. proposed turbidity for the optimal regression (L2 penalty parameter of 4)