Journal of Environmental Statistics

September 2015, Volume 7, Issue 5. http://www.jenvstat.org

Regression of Algae Biomass over Variables with Disjoint Spatial Support

Kevin Nichols
Department of Mathematics
Cal State Fullerton

Elena Trevino
Department of Mathematics
Cal State Fullerton

Calvin Pham
Southern California Coastal
Water Research Project

Jane Giamporcaro
Department of Mathematics
Cal State Fullerton

Anthony Rambone
Department of Mathematics
Cal State Fullerton

Atousa Karimi
Department of Mathematics
Cal State Fullerton

Kali Chowdhury
Department of Mathematics
Cal State Fullerton

Abstract

Algae biomass in California watersheds are impacted by several covariates including atmospheric nitrogen, watershed nutrients, land-use variables and physical-habitat variables. However, with several different agencies collecting and reporting data on both biomass response variables and subsets of all potential predictors, the result is several variables contained within multiple datasets, each compiled with data disjoint both in space and time from the other datasets. In this paper, the authors discuss a spatial statistical technique for forcasting values for each response and predictor over a shared spatial support and then using weighted standardized regression to identify which predictive variables are most important in explaining variability in algae biomass levels. Results will indicate that algae biomass levels are consistently correlated with the following: N03, N0x, Total Nitrogen and some land-use variables, while physical-habitat variables and atmospheric nitrogen are less successful predictors of algae biomass population density.

Keywords: Kernel Smoothing, Kriging, Shared Spatial Support, Standardized Regression, Descriptive Models, Algae Biomass, Atmospheric Nitrogen.

1. Introduction

California's watershed constituents, namely, rivers and streams, are a critical natural source to sustain indigenous aquatic and terrestrial natural species. Pollution of streams caused by urbanization has resulted in increased loading of contaminants to streams not only from direct infusion into water bodies but also by wet and dry atmospheric deposition. The State of California is proceeding with development of specific watershed management goals to define nutrient water quality objectives that may ultimately become environmental policies and regulations. A key element to be considered in development of these objectives (and eventual sound policies) is the level of nitrogen from natural vs. anthropogenic sources and their effect on biomass, which is characterized by three response variables of interest. The purpose of this project is to provide critically needed insight into the problem of nitrogen pollution, not only so that regulators can set appropriate limits, but also so that 'dischargers,' (the anthropogenic sources) can understand to what extent they affect, and can therefore control, undesirable levels of nitrogen contamination.

There is a general lack of understanding regarding natural vs. anthropogenically-influenced rates of atmospheric nitrogen deposition, due to the difficulty in separating the contribution of each variables to the overall level of nitrogen deposition. To further complicate this problem is the fact that several variables of interest are maintained by separate agencies and are recorded at potentially disjoint spatial locations. With several potential confounding variables to consider in addition to atmospheric nitrogen, such as land-use variables, physical-habitat variables and nutrient variables, it is difficult using traditional statistical techniques for estimating which variables have the largest impacts on algae biomass within a watershed.

Our proposed approach to the problem is to employ a statistical protocol for identifying which of several potential predictors, contained within spatially-disjoint geo-statistical datasets, most impacts algae biomass. Our proposed methodology combines protocols for addressing incomplete data, for spatial kernel smoothing of several datasets over a shared discretized or 'pixelated,' spatial support, for Kriging smoothed values, for determining the weight at each pixel over our support and finally an adaptation to weighted least squares regression with standardized coefficients of all the predicted variables to algae biomass response variables on a per pixel basis.

2. Data

Three bodies of data were provided by the Southern California Coastal Water Research Project (SCCWRP) . The first, henceforth the SCCWRP data, includes data from 1264 sites over nine ecological regions (Central Lahotan, Central Valley, Coastal Chapparal, Deserts Modoc, Interior Chaparral, North Coast, South Coast Mountain, South Coast Xeric and West Sierra) and among the 194 variables included for each site, contains measurements of nutrient concentrations, algal biomass and site-specific land-use and physical-habitat factors. This data is temporally supported from June 5, 2007 through May 31, 2012. Each observation contains aggregate measures of nutrient concentrations and algae biomass over time as well as values for the site-specific land-use and physical-habitat factors as were measured at the end of the temporal domain. Missing data frequently appeared throughout various fields of this data set.

The second set of data was constructed by the Community Air Quality Model (CMAQ). This

dataset contained 6,012 temporally summarized measures of wet, dry and total atmospheric nitrogen deposits. Each observation from this dataset is an aggregate summary of wet, dry and total atmospheric nitrogen from 2002-2006 as well as from 2004-2008.

The third and final dataset was collected by the National Atmospheric Deposition program (NADP) and has the same spatial support as the CMAQ data. Contained within the NADP data is 6,012 temporally-aggregate summaries of wet, dry and total atmospheric nitrogen deposits from 2002-2006, 2004-2008 and from 2007-2011. Both the CMAQ and NADP datasets are complete datasets. The NADP dataset is limited in that it potential does not estimate deposition in high-elevation areas well (Latysh and Weatherbee 2012).

It is important to note that the only time period in which all three datasets are temporally supported is from June 2007 through December 2008. However, the temporal resolution of our data has made isolating this particular period within the data impossible. Our conclusions will ultimately be somewhat restricted by the fact that we are using predictor variables from the SCCWRP data (summarized over 2007-2011) to estimate our response variables from the CMAQ and NADP data (summarized from 2004-2008). We are encouraged by researchers at SCCWRP that maintain that nutrient concentrations, site-specific land-use and physical-habitat factors may change significantly over very long periods of time but that using summarized data from 2007-2011 as a proxy for the same information from 2004-2008 is not overly irresponsible.

Each of the 1264 rows of measurements in the SCCWRP database had measurements on 174 variables. This set of data was initially assessed to propose which variables to delete and which to keep in an analysis dataset. Missing data accounted for 20.5% (n=50,253) of the 245,895 data values, and 49 of the 194 variables contained missing data on 30% or more of the observed values. Twenty-one scaled variables contained truncated data. Considerations for multi-collinearity and discussions with our ecology experts at SCCWRP resulted in the deletion of a total of 146 variables, resulting in a final set of k=48 potential predictors for the three potential responses from the SCCWRP database to be used for analysis. Appendix 1 provides a summary of variables retained for analysis.

Imputation using the missForest package in R was implemented to address the missing data problems inherent within the SCCWRP dataset. This iterative imputation method uses a random forest, or statistical algorithm to cluster points of data in functional groups, by fitting the observed data to predict the missing data until the algorithm has reached a predetermined stopping criterion or maximum number of iterations (Stekhoven and Buehlmann 2012).

3. Methods

The goal of this process is to identify which among our 48-predictors are having the largest impact/correlation with algae biomass response variables. The working assumption for this procedure is that the geo-statistical variables can be modeled as follows:

$$Z(s) = \mu(s) + f(\gamma(h)) \tag{1}$$

where $\mu(s)$ is spatially deterministic and in-homogenous and $\gamma_(h)$ is the semi-variogram of the data value once the deterministic aspect has been removed (i.e. semi-variogram for $Z(s) - \mu(s)$). The underlying assumption here is that once the deterministic aspect of the data has

been removed, the resulting data reflects an intrinsically stationary stochastic process from which we can forecast to new locations using ordinary Kriging with a parametric spherical variogram (Matheron 1963, Cressie 1990, Cressie 1993).

We begin by 'pixelating,' our data over a shared spatial support. This involves making ad hoc choice of a kernel $K(\mathbf{H})$ for spatial kernel smoothing over a region of California that is supported by all three datasets and an ad hoc choice for the number of pixels to regress over. Literature will support that often the choice of kernel makes little difference in results, but that choice of bandwidth is critical to avoid over/under smoothing, see for example Helmstetter et al. (2006).

Our ad hoc choice of spatial kernel is the symmetric Gaussian,

$$K(\mathbf{H}, s_1, s_2) = \frac{1}{2\pi \mathbf{H^2}} e^{\frac{\Delta(s_1,s_2)^2}{2\mathbf{H^2}}} \tag{2}$$

where $\mathbf{H}$ is a symmetric bandwidth selected using cross validated maximum likelihood estimation, see Helmstetter et al. (2006). Our ad hoc choice for the size and number of pixels to disjointly cover the region of California over which we have reasonable data coverage was 3,735 (.137 by .063, Longitude by Latitude) discretized pixels.

At any given location s_0, kernel smoothed estimates $\hat{\mu}(s_0)$ were computed based on our kernel and the spatial locations, $s_1, s_2, ..., s_n$ of the observed geo-statistical data points $Z(s_1), Z(s_2)..., Z(s_n)$ as follows:

$$\hat{\mu}(s_0) = \frac{\sum_{i=1}^{n} Z(s_i) K(\mathbf{H}, \mathbf{s_i}, \mathbf{s_0})}{\sum_{i=1}^{n} K(\mathbf{H}, \mathbf{s_i}, \mathbf{s_0})} \tag{3}$$

Plots or 'Heat maps,' of the kernel regressed predictors and responses are easy to construct at this point and serve as helpful tools in verifying which of the predictors are spatially most correlated with each of the response variables. Once the deterministic portion of our model $\hat{\mu}(s_0)$ is assessed at location s_0, values for $\hat{Z}(s_0) - \hat{\mu}(s_0)$ can be imputed using a Kriging procedure. The resulting sum of the deterministic and stochastic forecasts at s_0 is $\hat{Z}(s_0)$.

At any given location s_0, kernel density estimates were computed as follows:

$$\hat{f}(s_0) = \sum_{i=1}^{n} K(\mathbf{H}, \mathbf{s_i}, \mathbf{s_0}) \tag{4}$$

While a kernel density is not a meaningful tool for geo-statistical data as the locations at which data collected are not random, a kernel density in this instance is helpful in identifying how much data support there is local to each pixel. At a given pixel location, s_0, let us consider the following collections of spatially kernel smoothed and Kriged response values: $(\hat{Z}_{y1^*}(s_0), \hat{Z}_{y2^*}(s_0), \hat{Z}_{y3^*}(s_0))$ and spatially kernel smoothed and Kriged predictors $(\hat{Z}_{x1^*}(s_0), ..., \hat{Z}_{xk^*}(s_0))$. Note that x_j^* is predictor x_j in standardized form.

In addition to kernel smoothed and Kriged responses and predictors at each pixel s_0, kernel density estimates for standardized responses and standardized predictors were computed at

each pixel: $\hat{f_{y_1^*}}(s_0), \hat{f_{y_2^*}}(s_0), \hat{f_{y_3^*}}(s_0), \hat{f_{x_1^*}}(s_0)...\hat{f_{x_k^*}}(s_0)$. Weights $w(s)$ for weighted least squares regression at each pixel were computed as follows:

$$w(s_0) = \frac{\hat{f_{x_i^*}}(s_0) + \hat{f_{x_j^*}}(s_0)}{max(\hat{f_{x_i^*}}) + max(\hat{f_{x_j^*}})} \tag{5}$$

where x_i^* and x_j^* are any variable from the SCCWRP and NADP/CMAQ databases respectively.

Using the kernel regressed values for each of the three response variables, the kernel regressed values for all of the potential standardized predictive variables and the weights, three standardized weighted least squares regression models were proposed. The impact of each of the predictors on the response was determined by ranking the standardized coefficients of the model.

4. Results

Figure 1 contains heat maps of the three algal biomass response variables, PCTMAP, CHLA and AFDM. Of interest is the effect of atmospheric nitrogen deposition as was measured by the following three predictor variables, CMAQ Interpolated Wet Nitrogen from 2004-2008, CMAQ Interpolated Dry Nitrogen from 2004-2008 and CMAQ Interpolated Total Nitrogen from 2004-2008. Figure 2 contains heat maps of the NO3 and NOx from 2007-2011, two of the most effective predictors of the three responses while Figure 3 contains heat maps of the CMAQ Interpolated Nitrogen variables from 2004-2008.

Based on comparison of standardized slopes for our final weighted least squares regression models, NO3, NOx and Total Nitrogen were amongst the most impacting predictors in all three models. Other high impact predictors included land use variables like road density and percent land use catchment. Notice how spatially similar NO3 and NOx are with the three algal biomass response variables. It is not surprising that they are the dominant terms in the standardized regression models.

5. Discussion

While the results of this paper merely verify what ecologists have long known, namely that nitrogen based nutrients like NH4, NH3, NH2, NOx and Total Nitrogen are algal biomass population limiting nutrients, the weakly correlated relationship between atmospheric nitrogen and algal biomass is somewhat surprising. A potential explanation for this might simply be that atmospheric nitrogen from 2004-2008 contributes to nitrogen levels in local streams and lakes which are represented in the SCCWRP watershed nutrient data. The result being that watershed nitrogen levels are population density limiting variables and have a direct impact on algae biomass while atmospheric nitrogen has more of an indirect relationship with algae biomass.

This research implemented a new descriptive methodology for comparison of geo-statistical variables with disjoint spatial supports. It should be emphasized that even after Kriging there are still pockets of spatial non-stationarity which can severely deflate standard errors

for regression coefficients. Any attempt at prediction or statistical inference is ill-advised. However, descriptively, standardized regression coefficients of kernel smoothed and Kriged geo-statistical variables are a valuable new tool in identifying important predictors of responses for data with spatially disjoint supports.

References

Cressie NA. 1993. *Statistics for Spatial Data (Revised Edition).* Wiley: New Jersey.

Cressie NA. 1990. The Origins of Kriging. *Mathematical Geology* **22**: 239-252.

Helmstetter A, Kagan Y, Jackson D. 2006. Comparison of short-term and time-independent earthquake forecast models for southern California, *Bulletin of the Seismological Society of America* **96**, 90-106.

Latysh NE, Wetherbee GA. 2012. Improved mapping of National Atmospheric Deposition Program wet-deposition in complex terrain using PRISM-gridded data sets. USGS Staff - Published Research. Paper 736.

Matheron G. 1963. Principals in geostatistics. *Economic Geology* **58**: 1246-1266.

Stekhoven DJ, Buehlmann P. 2012. MissForest - nonparametric missing value imputation for mixed-type data. *Bioinformatics;***28(1)**: 112-118.

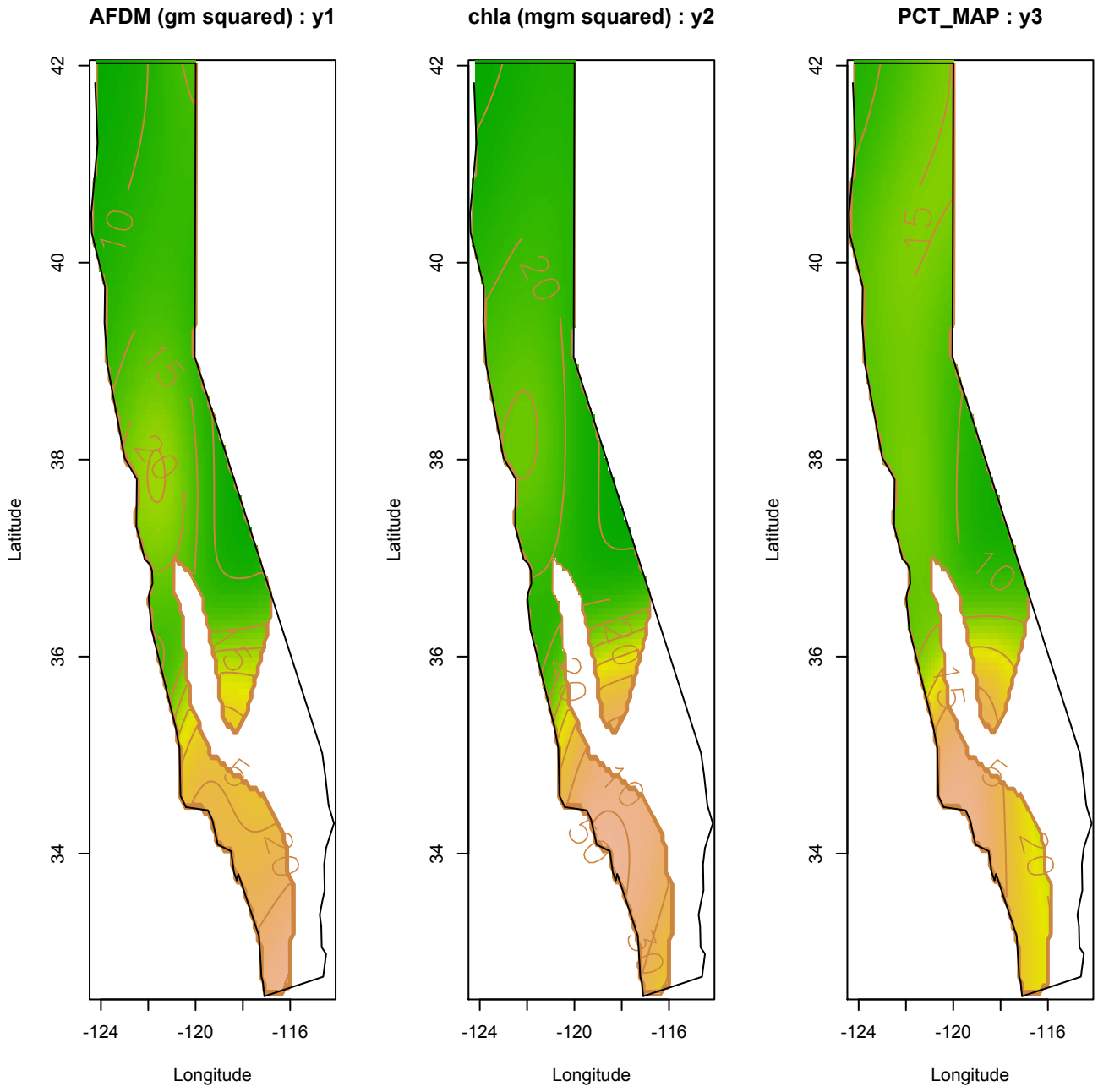

Figure 1: Results of spatial kernel regression of three algae biomass response variables based on 1,152 observations.

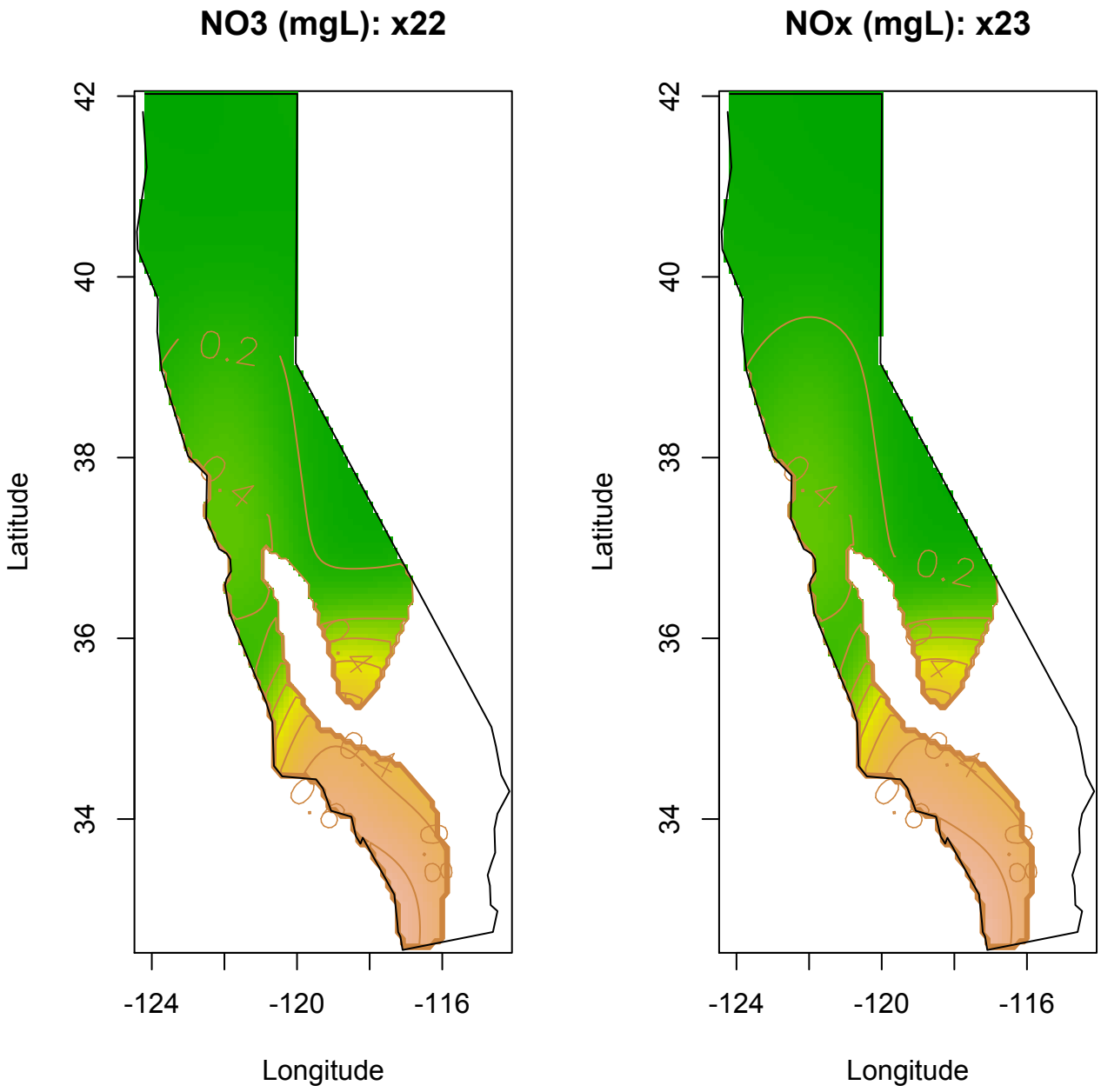

Figure 2: Results of spatial kernel regression of NO_3 and NO_x based on 1,152 observations.

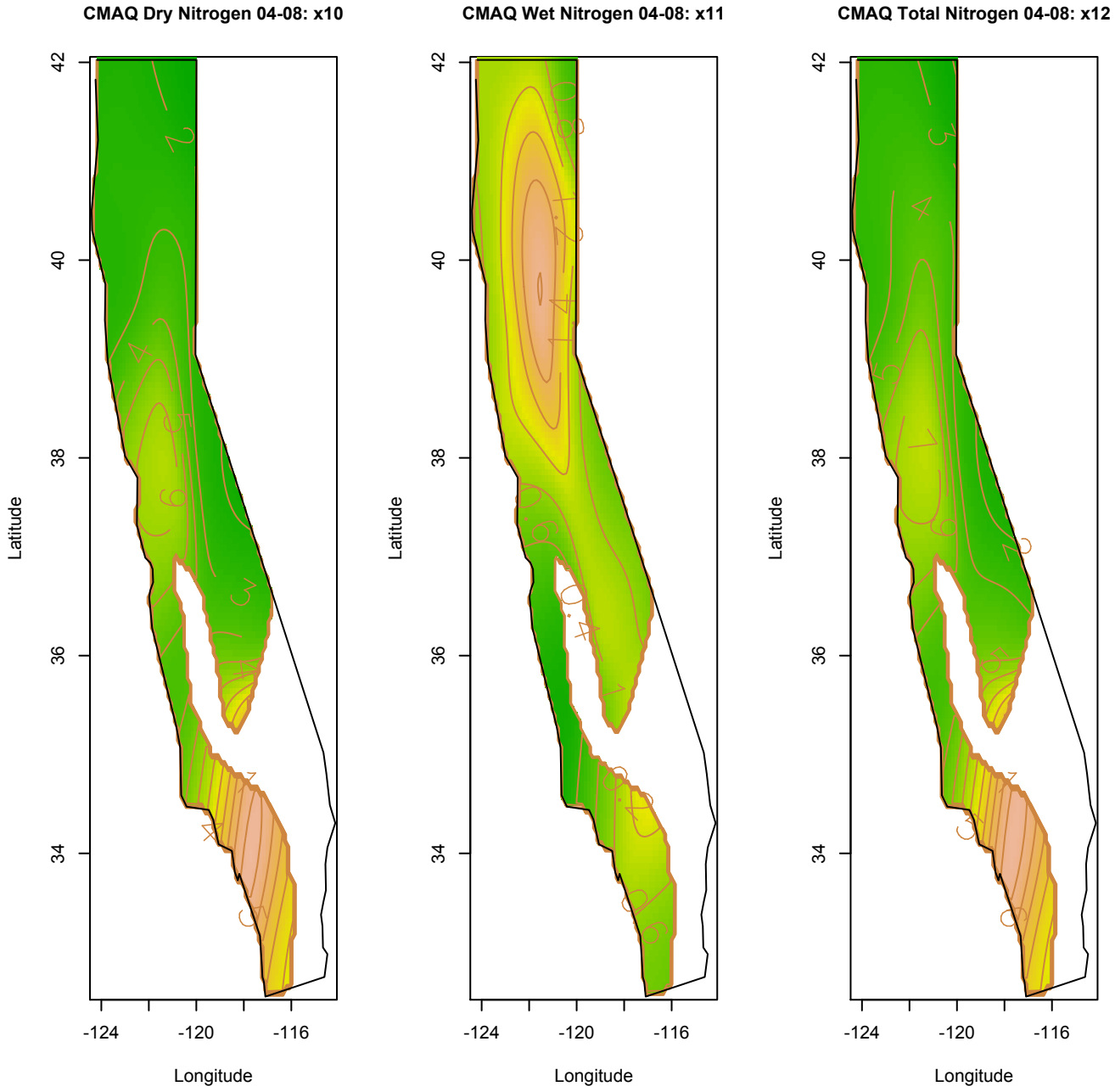

Figure 3: Results of spatial kernel smoothing of CMAQ interpolated nitrogen based on 6,022 observations.

Appendix 1: Variable Descriptions

Description	**Short Hand Expression**	**Regression Coefficient**
Ash-free dry mass corresponding to organic content in the sample	AFDM_gm2	y_1
Percent agricultural land use in catchment within 1-km radius from sampling site	Ag_2000_1K	x_1
Percent agricultural land use in catchment within 5-km radius from sampling site	Ag_2000_5K	x_2
Percent agricultural land use in catchment	Ag_2000_WS	x_3
Alkalinity	Alkalinity_mgL	x_4
Chlorophyll-a (photosynthesis pigment)	chla_mgm2	y_2
Chloride	Chloride_mgL	x_5
Percent land use in catchment within 1-km radius from sampling site	CODE_21_2000_1K	x_6
Percent land use in catchment within 5-km radius from sampling site	CODE_21_2000_5K	x_7
Percent land use in catchment	CODE_21_2000_WS	x_8
Conductivity	Conductivity_uScm	x_9
CMAQ Interpolated Dry Nitrogen 2004-2008	CMDN0408kghayr	x_{10}
CMAQ Interpolated Wet Nitrogen 2004-2008	CMDW0408kghayr	x_{11}
Flow Discharge (Q) (metric)_m3/s	FL_Q_M	x_{12}
Inverse distance (km) to nearest upstream dam in catchment	InvDamDist	x_{13}
Index of riparian disturbance, observational data that tallies all the different human impact	MaxOfW1_HALL	x_{14}
Mines within a 5-km radius	MINES_5K	x_{15}
Mines within the watershed radius	MINES_WS	x_{16}
mean monthly solar radiation (same month)	Mmsolar_sameMonth	x_{17}
NADP Interpolated Wet Nitrogen 2002-2010	NWN0210kghayr	x_{18}
NADP Interpolated Wet Nitrogen 2007-2011	NWN0711kghayr	x_{19}
Ammonium	NH4_mgL	x_{20}
Nitrogen + Oxygen2	N02	x_{21}
Nitrogen + Oxygen3	N03	x_{22}
Nitrate + Nitrite	NOx_mgL	x_{23}
paved intersections within a 1-km radius	PAVED_INT_1K	x_{24}
paved intersections within a 5-km radius	PAVED_INT_5K	x_{25}
paved intersections within a watershed radius	PAVED_INT_WS	x_{26}
Percent cover of coarse particulate organic matter in streambed	PCT_CPOM	x_{27}
Percent cover in fine substrata in streambed	PCT_FN	x_{28}
Macroalgal percent cover	PCT_MAP	y_3
Percent that was not sediment	PCT_NOSED	x_{29}
Percent sand + fines in streambed	PCT_SAFN	x_{30}

Percent sediment	PCT_SEDIM	x_{31}
Road density within a 1-km radius	RoadDens_1K	x_{32}
Road density within a 5-km radius	RoadDens_5K	x_{33}
Road density within the Watershed radius	RoadDens_WS	x_{34}
Slope	Slope_0	x_{35}
Mean of the slope within a 1-km radius	slopePercMean_1K	x_{36}
Mean of the slope within a 5-km radius	slopePercMean_5K	x_{37}
Mean of the slope within the watershed radius	slopePercMean_ws	x_{38}
Temperature	Temperature_C	x_{39}
Total nitrogen	TN_mgL_CALC	x_{40}
Total phosphorus	TP_mgL	x_{41}
Percent urban land use in catchment within a 1-km radius from sampling site	URBAN_2000_1K	x_{42}
Percent urban land use in catchment within a 5-km radius from sampling site	URBAN_2000_5K	x_{43}
Percent urban land use in catchment	URBAN_2000_WS	x_{44}
Percent canopy cover	XCDENMID	x_{45}
Slope	XSLOPE	x_{46}
Depth of stream	XWDEPTH	x_{47}
Width by depth	XwidthXdepth	x_{48}

Appendix 2: Regression Coefficients

Std. Predictor \ Std. Response	$\hat{Z}_{y1}(s_0)$	$\hat{Z}_{y2}(s_0)$	$\hat{Z}_{y3}(s_0)$
$\hat{Z}_{x1^*}(s_0)$	0.11	0.04	-0.03
$\hat{Z}_{x2^*}(s_0)$	-0.23	-0.06	0.12
$\hat{Z}_{x3^*}(s_0)$	0.01	-0.10	-0.65
$\hat{Z}_{x4^*}(s_0)$	0.10	-0.18	0.20
$\hat{Z}_{x5^*}(s_0)$	-0.21	-0.04	-0.39
$\hat{Z}_{x6^*}(s_0)$	-0.21	-0.01	-0.04
$\hat{Z}_{x7^*}(s_0)$	**0.68**	-0.08	-0.19
$\hat{Z}_{x8^*}(s_0)$	0.46	**0.67**	0.38
$\hat{Z}_{x9^*}(s_0)$	0.42	0.49	**1.10**
$\hat{Z}_{x10^*}(s_0)$	0.07	0.05	-0.01
$\hat{Z}_{x11^*}(s_0)$	0.00	0.01	0.05
$\hat{Z}_{x12^*}(s_0)$	-0.09	-0.17	0.21
$\hat{Z}_{x13^*}(s_0)$	0.20	0.13	0.00
$\hat{Z}_{x14^*}(s_0)$	0.26	-0.20	-0.44
$\hat{Z}_{x15^*}(s_0)$	0.04	-0.06	0.11
$\hat{Z}_{x16^*}(s_0)$	-0.09	0.08	0.10
$\hat{Z}_{x17^*}(s_0)$	-0.09	0.02	-0.07
$\hat{Z}_{x18^*}(s_0)$	0.00	0.03	0.10
$\hat{Z}_{x19^*}(s_0)$	-0.01	-0.06	-0.13
$\hat{Z}_{x20^*}(s_0)$	-0.08	-0.39	-0.04
$\hat{Z}_{x21^*}(s_0)$	-0.07	0.17	-0.09
$\hat{Z}_{x22^*}(s_0)$	**-1.05**	**-0.64**	-0.39
$\hat{Z}_{x23^*}(s_0)$	**1.04**	0.24	**1.27**
$\hat{Z}_{x24^*}(s_0)$	-0.32	-0.57	0.04
$\hat{Z}_{x25^*}(s_0)$	-0.16	0.05	-0.38
$\hat{Z}_{x26^*}(s_0)$	-0.08	0.01	-0.23
$\hat{Z}_{x27^*}(s_0)$	0.05	-0.08	0.30
$\hat{Z}_{x28^*}(s_0)$	0.17	-0.02	0.13
$\hat{Z}_{x29^*}(s_0)$	0.36	0.26	0.02
$\hat{Z}_{x30^*}(s_0)$	-0.18	-0.03	-0.21
$\hat{Z}_{x31^*}(s_0)$	0.00	-0.06	-0.02
$\hat{Z}_{x32^*}(s_0)$	**-0.72**	0.11	0.14
$\hat{Z}_{x33^*}(s_0)$	**-0.62**	**-1.00**	-0.22
$\hat{Z}_{x34^*}(s_0)$	**0.58**	0.45	0.12
$\hat{Z}_{x35^*}(s_0)$	0.06	0.10	0.02
$\hat{Z}_{x36^*}(s_0)$	-0.02	0.04	-0.02
$\hat{Z}_{x37^*}(s_0)$	0.17	0.13	-0.32
$\hat{Z}_{x38^*}(s_0)$	-0.01	0.11	0.45
$\hat{Z}_{x39^*}(s_0)$	0.22	0.22	0.23
$\hat{Z}_{x40^*}(s_0)$	**0.65**	**0.65**	**-1.25**

$\hat{Z}_{x41^*}(s_0)$	-0.11	0.15	0.40
$\hat{Z}_{x42^*}(s_0)$	**0.66**	**0.57**	**0.70**
$\hat{Z}_{x43^*}(s_0)$	-0.29	-0.48	-0.46
$\hat{Z}_{x44^*}(s_0)$	-0.13	0.40	0.43
$\hat{Z}_{x45^*}(s_0)$	-0.04	-0.06	-0.07
$\hat{Z}_{x46^*}(s_0)$	0.14	-0.18	-0.30
$\hat{Z}_{x47^*}(s_0)$	-0.06	-0.30	0.19
$\hat{Z}_{x48^*}(s_0)$	0.06	0.29	-0.32

Affiliation:

Kevin Nichols
McCarthy Hall 154, Department of Mathematics, California State University
800 N. State College Blvd., Fullerton, CA 92831
Telephone: 657-278-3631
Fax: 657-278-3972
E-mail: knichols@fullerton.edu