International Journal of Geosciences, 2014, 5, 1609-1617
Published Online December 2014 in SciRes. http://www.scirp.org/journal/ijg
http://dx.doi.org/10.4236/ijg.2014.513132

Scientific Research

Multivariate Geostatistical Model for Groundwater Constituents in Texas

Faye Anderson

Texas A & M AgriLife Research Center at Beaumont, College Station, Texas, USA
Email: andersonfaye7@gmail.com

Received 2 October 2014; revised 28 October 2014; accepted 18 November 2014

Abstract

Although many studies have explored the quality of Texas groundwater, very few have investigated the concurrent distributions of more than one pollutant, which provides insight on the temporal and spatial behavior of constituents within and between aquifers. The purpose of this research is to study the multivariate spatial patterns of seven health-related Texas groundwater constituents, which are calcium (Ca), chloride (Cl), nitrate (NO_3), sodium (Na), magnesium (Mg), sulfate (SO_4), and potassium (K). Data is extracted from Texas Water Development Board's database including nine years: 2000 through 2008. A multivariate geostatistical model was developed to examine the interactions between the constituents. The model had seven dependent variables—one for each of the constituents, and five independent variables: altitude, latitude, longitude, major aquifer and water level. Exploratory analyses show that the data has no temporal patterns, but hold spatial patterns as well as intrinsic correlation. The intrinsic correlation allowed for the use of a Kronecker form for the covariance matrix. The model was validated with a split-sample. Estimates of iteratively re-weighted generalized least squares converged after four iterations. Matern covariance function estimates are zero nugget, practical range is 44 miles, 0.8340 variance and kappa was fixed at 2. To show that our assumptions are reasonable and the choice of the model is appropriate, we perform residual validation and universal kriging. Moreover, prediction maps for the seven constituents are estimated from new locations data. The results point to an alarmingly increasing levels of these constituents' concentrations, which calls for more intensive monitoring and groundwater management.

Keywords

Multivariate, Geostatistical, Groundwater, Constituents, Texas

1. Introduction

Groundwater is one of Texas natural resources that supplies the majority of the total water use in Texas [1].

Many local groundwater management districts are organized on county lines rather than on natural boundaries of the aquifers. Few counties take their water from more than one aquifer [2]. Groundwater quality covers physical, chemical, and biological aspects. Physical water qualities include temperature, turbidity, color, taste, and odor [3]. The increasing demand for high quality groundwater has driven many studies to investigate how the constituents' concentrations change over time and space, and what sources can be controlled in order to keep their levels within the acceptable range. Higher concentrations negatively affect the environment and public health. Examples of groundwater pollution sources include salt water intrusion, fertilizer leakage, natural erosions, and mine discharge.

This study focuses on the simultaneous spatial and temporal distributions of the seven most investigated groundwater constituents (five major and two minor constituents) over a nine-year period. A deeper understanding of the factors affecting constituents' levels may lead to a more successful and specialized programs designed at protecting groundwater from contamination. The results of this research project may be relevant to preventing and controlling groundwater contamination.

2. Methods

2.1. Data

Data was obtained from the Texas Water Development Board Groundwater Database for all Texas wells from 2000 through 2008. The samples were checked for flagged values to ensure acceptable results in terms of reliable sampling, threshold conditions, or other criteria that can label the values as non-reliable. The wells are sampled periodically every four years (**Figure 1**). R version 2.9.0 was used. Descriptive statistics were performed using the build-in functions within the stats package. Variograms were plotted using geoR package.

2.2. Investigation of Temporal and Spatial Effects

To investigate temporal effects, Fisher's F-test was used to test the null hypothesis of non-changing variances between samples of the following years: 2000 versus 2004, 2001 versus 2005, 2002 versus 2006, 2003 versus 2007 and 2004 versus 2008. Fisher's F test p-values were all greater than the significance level of 0.05. Hence we cannot reject the null hypothesis of equal variances. Therefore, none of the constituents has shown any change in variance from the year 2000 to the year 2008. Repeated measures t-test was used to test the null hypothesis of non-changing means for the seven constituents (calcium, chloride, magnesium, potassium, nitrate, sulfate, sodium). The test was run to compare sample pairs of the following years: 2000 versus 2004, 2001 versus 2005, 2002 versus 2006, 2003 versus 2007 and 2004 versus 2008. It was found that none of the constituents has shown any change from the year 2000 to the year 2008. T-tests p-values were all greater than the significance level of 0.05. Hence the null hypothesis of equal means could not be rejected. Moreover, mapping of annual concentrations showed that for each of the seven constituents, the differences between two years levels was around zero. Based on the results for non-changing variances and means across the years, it was concluded that the data set does not contain temporal effects and all records between 2000 and 2008 were combined.

2.3. Exploratory Data Analyses

Exploratory analyses showed that Texas well depth means are similar all over the State, and that variances of constituents are lesser within an aquifer than between the aquifers. This means that the locality of a constituent has an effect on its level, which calls for a spatial model. Descriptive statistics showed high shifts of skewness and kurtosis from 0 and 3, respectively, which are the characteristic values of normal distribution. Therefore, the variables generally exhibit non symmetric distributions, with long tails and several outliers (**Table 1**). Because the data also had positive skewness as well as outliers, a log transformation on all the constituent data was performed. Furthermore, standard tests of univariate and multivariate normality (Shapiro-Wilk test and E-statistic (Energy)) [4] did not reject the null hypothesis of normality for the log-transformed data. Four records out of 3379 contained zero values. We neglected these records. **Table 2** presents the pair-wise correlations between the transformed constituents across aquifers whereas **Table 3** lists correlations within the Ogallala Aquifer, which are all non-negligible. Please note that the absence or drop of correlation between constituents does not imply the absence of spatial correlation.

A preliminary variogram analysis enables us to visualize the characterization of spatial correlation. At this

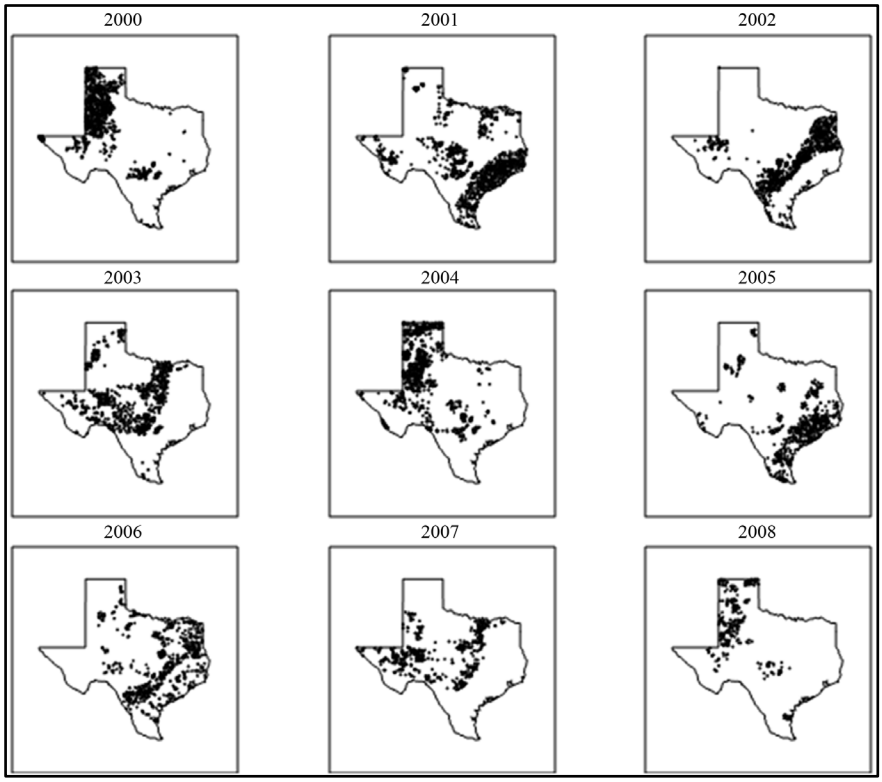

Figure 1. Groundwater four-year periodical sampling from 2000 to 2008.

Table 1. Across-aquifer descriptive statistics of the groundwater constituents.

Variable	Mean	Minimum	Maximum	Median	Std. Dev.	Skewness	Kurtosis
Calcium	93.72	0.35	1560.00	68.50	104.38	4.19	25.80
Chloride	159.32	1.85	17000.00	41.60	421.6111	18.41	635.51
Potassium	5.95	0.20	99.30	4.40	6.20	3.53	25.69
Magnesium	36.59	0.22	559.00	24.20	44.60	3.93	24.23
Sodium	122.51	2.39	13900.00	47.15	301.76	24.74	1052.38
Nitrate	17.33	0.00	425.88	8.01	31.68	5.03	35.69
Sulfate	192.40	0.00	5110.00	46.50	389.27	4.08	22.56

Table 2. Pair-wise inter-constituent correlations across aquifers.

Variables	ln(Ca)	ln(Mg)	ln(Na)	ln(K)	ln(SO_4)	ln(NO_3)	ln(Cl)
ln(Ca)	1	0.695	0.043	0.218	0.403	0.428	0.359
ln(Mg)	0.695	1	0.166	0.525	0.557	0.489	0.315
ln(Na)	0.043	0.166	1	0.628	0.696	0.026	0.832
ln(K)	0.218	0.525	0.628	1	0.629	0.206	0.541
ln(SO_4)	0.403	0.557	0.696	0.629	1	0.288	0.678
ln(NO_3)	0.428	0.489	0.026	0.206	0.288	1	0.157
ln(Cl)	0.359	0.315	0.832	0.541	0.678	0.157	1

Table 3. Pair-wise inter-constituent correlations within the Ogallala Aquifer.

	Ca	Mg	Na	K	SO_4	NO_3	Cl
Ca	1.00	0.67	0.64	0.44	0.76	0.52	0.81
Mg		1.00	0.72	0.77	0.87	0.46	0.82
Na			1.00	0.59	0.86	0.41	0.88
K				1.00	0.64	0.43	0.61
SO_4					1.00	0.45	0.81
NO_3						1.00	0.46
Cl							1.00

point, we recognize that there may be mean effects that are unaccounted for. Nevertheless, these initial variogram estimates indicate consistent spatial behavior across aquifer. None of the constituents showed any significant anisotropy within an aquifer. Multivariate intrinsic correlation exists when the multivariate correlation of a multivariate data set is independent of the spatial correlation. Multivariate intrinsic correlation allows one to simplify data modeling. Under intrinsic correlation, the joint covariance matrix is given by the Kronecker product $\Sigma = V \otimes R$, where V is the standard variance-covariance matrix. We have $\Sigma^{-1} = V^{-1} \otimes R^{-1}$, which only requires the inversion of the 7×7 variance-covariance matrix V and the $n \times n$ spatial correlations matrix W [5]. One of the ways to test for intrinsic correlation is to calculate codispersion coefficients (*i.e*. the ratio of the cross versus the direct variograms). If the codispersion coefficients are constant, we can assume intrinsic correlation [5] [6]. The calculations of codispersion coefficients show that they are practically constant for all spatial scales. Their variances ranged from 0.01 to 0.04. This supports the intrinsic correlation hypothesis, *i.e*. a similar correlation matrix W holds at all spatial scales. In other words, when the covariance is structured according to the intrinsic correlation model, all direct and cross covariance functions are proportional to one basic covariance function. The model of intrinsic correlation is entirely specified by its spatial structure, and by the variance-covariance matrix [6].

2.4. Principal Component Analysis

Principal components analysis (PCA) was conducted to study and visualize the correlations between the variables and hopefully be able to limit the number of variables to be measured afterwards, and to visualize observations in a 2- or 3-dimensional space in order to identify similarities and dissimilarities within observations. PCA was performed using the correlation matrix, which brings the measurements onto a common scale. In other words, a constituent which concentration varies between 0 and 1 will not weigh more in the projection than a constituent varying between 0 and 400. The Scree Plot of **Figure 2** shows the cumulative variance explained by the seven principal components F1:F7. The principal components are sorted in decreasing order of variance, so the most important principal component is always listed first. Positions of the observations on a biplot are scores of the observations on the first two components. Distance between observations on this plane is "how close these observations are to each other" (Since we use two-dimensional plot we have ignored all remaining components). With respect to the first two principal components, nitrate is a bit isolated as compared to the other constituents. This suggests that nitrate levels might come from a different origin than the remaining constituents. The variation explained by each principal component does not exceed that of the first principal component which is 52%. The first two principal components explain 75% of the data variation. If further analysis needs to be performed on the 2-dimensional PCA scores instead of the seven-dimensional dataset, this leaves a significant amount of variation not accounted for. Therefore, principal components analysis results were not used to model the data.

3. Results and Discussion

For our estimation we used the K-Bessel (Matern) model for the semivariogram model since its smoothness can be adjusted. Although the computation are cumbersome, the advantage of this model is that the behavior of the

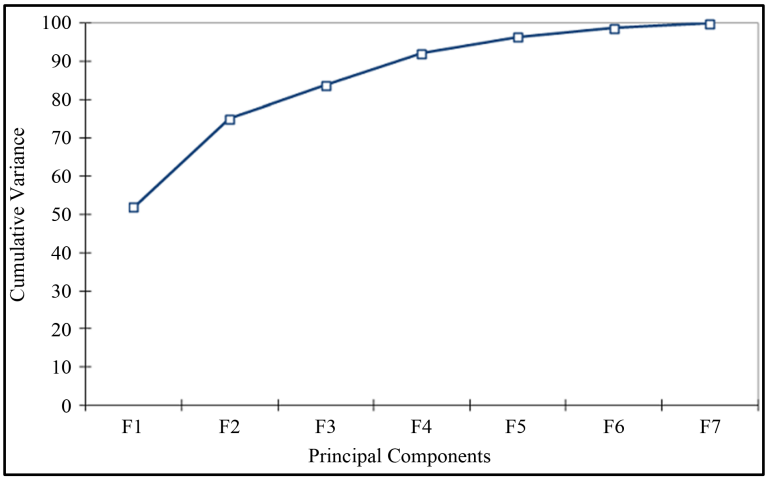

Figure 2. Scree plot.

semivariogram near the origin can be estimated from the data rather than assumed to be of a certain form. Also, changing the value of α, we can get other semivariogram models. For example, when $\alpha = 0.5$ we get the exponential covariance function [7]. We estimate β and $\Sigma(j)$ using iterative re-weighted generalized least squares IRWGLS [8]. For this model, the random functions are assumed to be second-order stationary. Isotropy of the data is also assumed. This means that the variogram is a function of the distance vector regardless of its direction. An alternative to IRWGLS would be maximum likelihood ML. We chose IRWGLS for edification purposes. The complexity of IRWGLS estimation lies in estimating the full covariance matrix of the data which is of a big size in our case. On the other hand, IRWGLS estimation is more efficient than weighted least squares and ordinary least squares. IRWGLS estimator of theta is efficient and has as little bias as possible [8].

In order to determine the significant covariates to include in the model, we perform a model selection exercise. The research started with six covariates: altitude, latitude, longitude, major aquifer, water level and percentage of irrigated acres per county. The effect of percentage of irrigated acres per county is excluded from the analysis since it has shown no association with any of the constituents. Model selection was performed to decide on which of the remaining covariates to include in our model. The criteria for model selection is Akaike's information criterion AIC. One thousand records were selected based on simple random sampling without replacement. Many regressions were performed and their AIC values were compared. We started with no intercept. Then single predictor models, then two-predictor models and so on. The regression model with the least AIC value (3143.04) was found to be [Ca, Cl, NO_3, Na, Mg, SO_4, K] = longitude + altitude + water depth + aquifer effect (**Table 4**).

Figure 3 shows the empirical variogram of the OLS residuals. The seven variograms follow similar spatial behavior (range). **Figure 4** shows the normalized residuals. The first seven standardized individual empirical variograms in **Figure 4** look very similar to each other. The only difference between the variograms is the sill value. After running the IRWGLS algorithm for four iterations the results converge with tolerance less than 1e−05. **Table 5** and **Figure 5** show the Matern covariance function parameters after the fourth iteration. Practical range is about 44 miles, variance is 0.83, nugget is zero and smoothness parameter was fixed at 2. The range indicates that the maximal distance at which the constituents are spatially auto-correlated is 44 miles. Beyond 44 miles, the distance among wells does not affect the spatial structure of the data forming a sill of 0.83. At zero distance the variogram is zero. That is, there is no nugget effect. Comparing OLS residuals variance-covariance results showed that there is more correlation represented in the fourth iteration estimates, which is an indicator that IRWGLS is more efficient than OLS. To validate the model, one hundred locations were randomly selected from the original dataset. This dataset was not used in the estimation analysis. It was used to predict the constituent levels at these sites and study the behaviour of the residuals. Residuals were computed using the IRWGLS fourth iteration estimates. The scatter plots of the residuals do not follow a specific pattern. The residuals departure from the normality was negligible. Shapiro-Wilk test did not reject the null hypothesis of normality for the residuals. P-values were less than the significance level of 0.05. From the above discussion we conclude that the assumptions were reasonable and the choice of the model was appropriate.

Table 4. Regression models AIC values.

Regression	AIC
Nitrate = 1	3643.62
Nitrate = Altitude	3530.97
Nitrate = Latitude	3573.52
Nitrate = Longitude	3507.15
…	…
Nitrate = Water Depth	3510.61
Nitrate = Longitude + Altitude	3509.11
Nitrate = Longitude + Latitude	3502.37
Nitrate = Longitude + Water Depth	3401.39
…	…
Nitrate = Altitude + Latitude + Longitude + Water Depth	3371.91
…	…

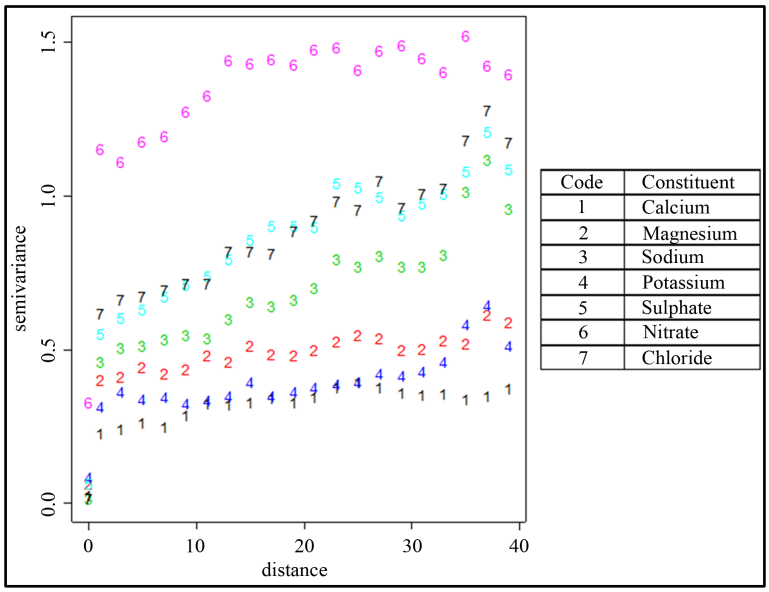

Figure 3. OLS residuals empirical variogram.

Because the concentrations were log transformed prior to modeling (**Table 6**), the predicted concentrations on the prediction maps are contrasted to the log transformed values of EPA's maximum contaminant levels [9] (**Figure 6**). Prediction maps provide an easier tool for researchers and hydrologists to see which constituents' concentrations might exceed or go below the EPA tolerated levels. This also helps groundwater policy makers to act appropriately and in a timely manner in order to protect our natural resources. A grid of two hundred and sixteen randomly chosen locations was formed. Elevations of points were estimated using global positioning system website GPSies. **Table 7** presents the constituents with predicted concentrations that exceed the EPA's maximum concentration levels.

4. Conclusion

This study focused on seven of the most researched groundwater constitutents in Texas. Namely, calcium (Ca),

Table 5. Matern covariance function parameters after fourth iteration, Kappa = 2.

Nugget (tausq) = 0
Range (miles) = 8.22
Sill (sigmasq) = 0.83
Kappa (smoothness parameter) = 2
Practical Range = 44.12 (miles)

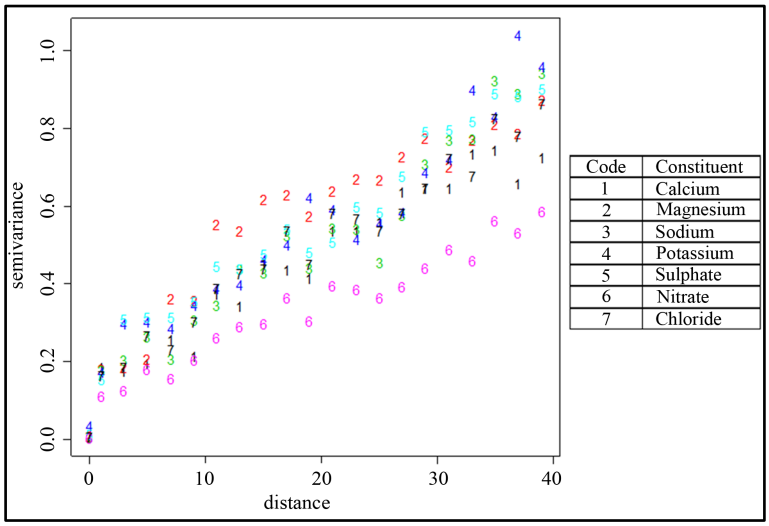

Figure 4. OLS standardized residuals empirical variogram.

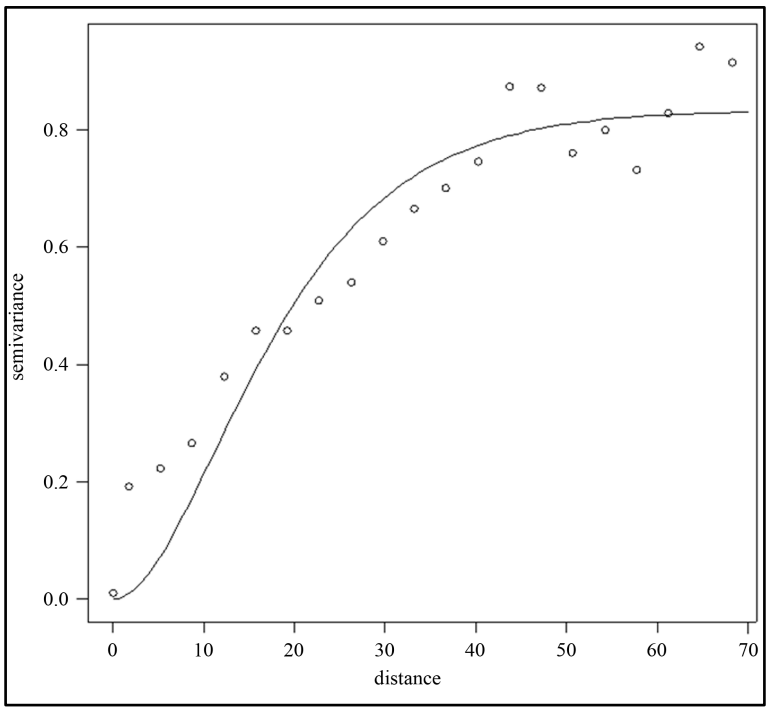

Figure 5. Fourth IRWGLS iteration matern function estimate variogram.

Table 6. EPA tolerated concentrations.

Constituent	Mx: The maximum level allowed in public water supplies by the EPA	Log(Mx)
Calcium	And magnesium 200 mg/L	5.298317
Chloride	250 mg/L	5.521461
Magnesium	And calcium 200 mg/L	5.298317
Nitrate	10 mg/L	2.302585
Potassium	10 mg/L	2.302585
Sodium	20 mg/L	2.995732
Sulfate	250 mg/L	5.521461

Table 7. Constituents with predicted concentrations higher than MCLs.

Constituent	Affected aquifers
Calcium	Gulf Coast, Ogallala, Edwards-Trinity Plateau and Seymour
Nitrate	Ogallala, Seymour and Edwards-Trinity Plateau
Chloride	Ogallala Edwards-Trinity Plateau and Hueco-Mesilla Bolson
Potassium	Ogallala
Sodium	All
Sulfate	Ogallala and Edwards-Trinity Plateau

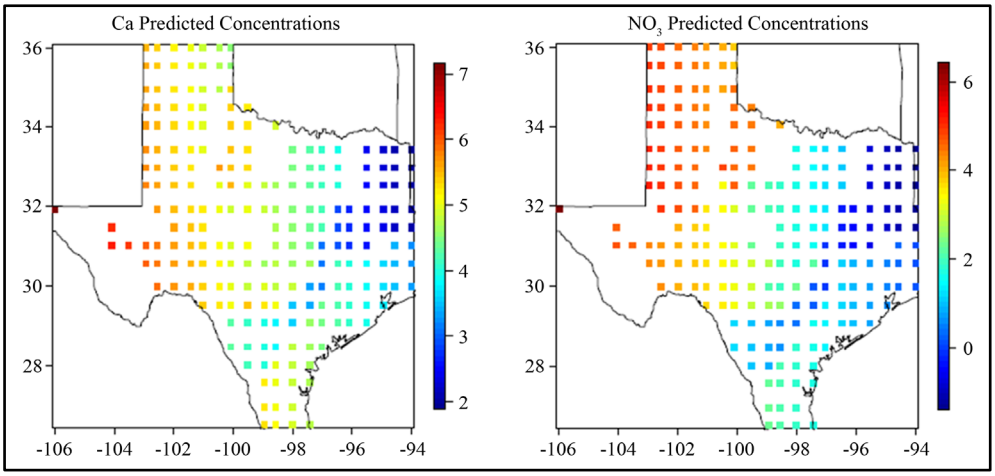

Figure 6. Predicted concentrations for five constituents in lnMg.

chloride (Cl), nitrate (NO_3), sodium (Na), magnesium (Mg), sulfate (SO_4), and potassium (K). Data were extracted from TWDB database for the nine years from 2000 to 2008. No temporal effects were found in the concentrations but data was found to be intrinsically correlated. This allowed to spatially model the between-constituent effects individually using IRWGLS. The multivariate geostatistical model used latitude, longitude, elevation, water depth and aquifer effect as the mean covariates. Also, a covariance structure was estimated to link the constituents' concentrations with the spatial structure. The non-temporally different concentrations of the seven constituents points to the value of geostatistical modeling techniques as a valuable estimation method. This study provides an example where simultaneous modeling of pollutants can be conducted based on locality, which closes a big gap in environmental research where interaction between variables is significant. This study also shows that calcium, chloride, nitrate, sodium, magnesium, sulfate, and potassium follow similar increasing

trends. The high concentrations that exceed EPA tolerated levels presented in the prediction maps (**Figure 6**) agree with recent research observations [10], which further validates the model as a tool for hydrologists and policy makers to accurately estimate the seven constituents simultaneously based on location.

References

[1] Texas Water Development Board (TWDB) (2007) Water for Texas 2007. Volume II, p. 161.

[2] Sansom, A., Armitano, E.R. and Wassenich, T. (2008) Water in Texas: An Introduction. UT Press.

[3] Westlake, K. (1995) Landfill Waste Pollution and Control. Horwood Publishing.

[4] Szekely, G.J. and Rizzo, M.L. (2013) Energy Statistics: A Class of Statistics Based on Distances. *Journal of Statistical Planning and Inference*, **143**, 1249-1272. http://dx.doi.org/10.1016/j.jspi.2013.03.018

[5] Piegorsch, W.W. and El-Sharaawi, A.H. (2002) Encyclopaedia of Environmentrics. John Wiley and Sons, Hoboken, 1368.

[6] Wackernagel, H. (2003) Multivariate Geostatistics: An Introduction with Applications. 3rd Edition, Springer. http://dx.doi.org/10.1007/978-3-662-05294-5

[7] Waller, L.A. and Gotway, C.A. (2004) Applied Spatial Statistics for Public Health Data. Wiley-IEEE.

[8] Schabenberger, O. and Gotway, C.A. (2005) Statistical Methods for Spatial Data Analysis. Chapman & Hall/CRC Press.

[9] United States Environmental Protection Agency (EPA) (2014) Drinking Water Contaminants. http://water.epa.gov/drink/contaminants/

[10] Chaudhuri, S. and Ale, S. (2014) Long Term (1960-2010) Trends in Groundwater Contamination and Salinization in the Ogallala Aquifer in Texas. *Journal of Hydrology*, **513**, 376-390. http://dx.doi.org/10.1016/j.jhydrol.2014.03.033

Scientific Research Publishing (SCIRP) is one of the largest Open Access journal publishers. It is currently publishing more than 200 open access, online, peer-reviewed journals covering a wide range of academic disciplines. SCIRP serves the worldwide academic communities and contributes to the progress and application of science with its publication.

Other selected journals from SCIRP are listed as below. Submit your manuscript to us via either submit@scirp.org or Online Submission Portal.