BSTT 562 Linear Model (PhD Level)

Final Project

Estimation for the unknown true concentration of environmental analytical data

University of Illinois at Chicago

School of Public Health

Division of Epidemiology & Biostatistics

Pei-Shan Yen

December 14 2018

Abstract

We use the statistical model to describe the environmental analytical measurements. To reduce the bias naturally occurring in the observation process, the measurements come from the interlaboratory data which is divided into three different levels of concentration respectively having five replications. The statistical model is constructed using the nonlinear format instead of the traditional linear format, for accounting for the larger variation of the analyte in a higher concentration level. We estimate the model parameters following the work of Bhaumik and Gibbons (2005), which adopts the method of moments that is fast in execution. After obtaining the model parameters, we further apply this model to estimate the unknown true concentration and verify the variation property of the analytic observations.

Contents

1.Introduction

(1) Research Background

Statistical methods have played a major role in the environmental researches. One specific topic is to reach an accurate estimation of the true concentration of groundwater pollution, since the traditional environmental analysis commonly accepts the original measurements as true concentration without considering the uncertainty, which may not be realistic. The characteristic of the analytical measurements is that the variations increases in higher concentration level. Hence, we adopt the nonlinear model proposed by Rocke and Lorenzato (1995) and is later improved by Bhaumik and Gibbons (2005) to account for the issue of uncertainty. The measurements come from the interlaboratory data. The model parameters using the method of moments as also suggested by Bhaumik and Gibbons (2005). After obtaining the model parameters, we proceed to applying this model to predict the true level of concentration. Finally, we perform the simulations to verify the estimation results.

(2) Dataset

We use the data from Bhaumik and Gibbons (2005) to present the estimation process of true concentration. The dataset contains the cadmium concentration from a blind interlaboratory study performed by the Ford Motor Company. Table 1 summarizes the details for the observations. These samples are divided into three groups having different concentration levels (0, 20, and 100μg/L), tested by five laboratories with five replications for each group.

Table 1. Interlaboratory Data for Cadmium (ug/L)

Lab	Replication	Concentration		
		0 ug/L	20 ug/L	100 ug/L
1	1	-3.000	10.000	92.000
	2	4.000	20.000	100.000
	3	-4.000	17.200	97.800
	4	3.000	24.000	100.000
	5	3.100	19.100	109.000
2	1	-0.060	17.815	90.455
	2	0.010	17.305	87.610
	3	0.115	16.570	85.550
	4	-0.055	17.360	89.925
	5	0.340	18.120	90.070
3	1	-7.400	27.100	107.400
	2	-2.100	19.400	108.100
	3	-11.400	9.000	83.800
	4	-11.100	10.500	81.900
	5	-1.400	19.300	94.200
4	1	1.000	21.000	96.000
	2	-2.126	16.049	90.650
	3	0.523	16.082	89.388
	4	-2.000	17.000	91.000
	5	-0.551	15.489	85.867
5	1	0.000	18.000	91.000
	2	0.000	19.000	101.000
	3	0.000	19.000	102.000
	4	-1.000	18.700	92.700
	5	0.038	19.790	99.884

* Data Set from Bhaumik (2005)

2. Methodology

(1) Model Selection

The environmental analyte exhibits a more substantial variation in the higher concentration level, which makes the traditional linear model, $y = \alpha + \beta x + e$, not appropriate to use because the linear model assumes a constant variation throughout the entire range of *x*. To resolve this issue, Rocke and Lorenzato (1995) propose a two-component nonlinear model, $y = \alpha + \beta x e^{\eta} + e$, where the nonlinear term e^{η} is added to model the proportionality between the measurement variation and the concentration. The model contains two error terms distributed following a normal distribution. With these two error terms, the constant variations at near-zero concentrations and the inflated variation at larger concentration can be both incorporated. In this report, we adopt an improved version of the nonlinear model proposed by Bhaumik and Gibbons (2005). Compared with the work of Rocke and Lorenzato (1995), this model is extended to incorporates interlaboratory measurements for reducing uncertainty.

$$y_{ijk} = \alpha_i + \beta_i x_j e^{\eta_{ijk}} + e_{ijk} \tag{1}$$

- y_{ijk} is the measurement
- Index i is the laboratory, where $i = 1, 2, ..., q$
- Index j is the level of concentration, where $j = 1, 2, ..., r$
- Index k is the order of the replicate measurement, where $k = 1, 2, ..., N_{ij}$
- η^{ijk} is the proportional error, and we assume $\eta^{ijk} \sim Normal(0, \sigma_{\eta}^2)$
- e^{ijk} is the additive error, and we assume $e^{ijk} \sim Normal(0, \sigma_e^2)$
- Assume η^{ijk} and e^{ijk} are independent

(2) Parameter Estimation

The method of moments is employed to estimate the model parameters (Bhaumik and Gibbons (2005)). The advantage of using the method of moments is twofold. First, it provides reasonable

estimates of the corresponding population moments. Second, it is asymptotically efficient which allows for a faster estimation stemming from the large-sample properties.

In conjunction with the method of the moments, we further use the data-separation technique to reach a better estimation of the model parameters. The dataset of low concentration level is applied to evaluate α_i and σ_e^2. On the other hand, high-level measurement is employed to assess β_i and σ_η^2.

In what follows, we will first describe the equations in computing the model parameters. Next, we present the formulas of the point estimate of unknown true concentration. Finally, we illustrate the approach to estimating the confidence interval of each concentration level.

(A) Estimation of α and σ_e^2

We use a total of twenty-five low concentration level measurement $y_{101}, y_{102}, \cdots, y_{505}$ among five laboratories to estimation α and σ_e^2. Then we determine both parameters using Eqs. (2) and (3)

$$\alpha_i = \frac{\sum_{k=1}^{5} y_{i0k}}{5} \ , \ i = 1,2,3,4,5 = \textit{the lab index} \ ; \ k = 1,2,3,4,5 = \textit{the replicate index} \qquad (2)$$

$$\sigma_e^2 = \frac{\sum_{i=1}^{5}\left(\sum_{k=1}^{5}\left(y_{i0k} - \bar{y}_{i0}\right)^2 / (5-1)\right)}{5} \qquad (3)$$

(B) Estimation of β and σ_η^2

We use a total of fifty high concentration level measurement $y_{111}, y_{112}, \cdots, y_{525}$ among five laboratories to estimation β and σ_η^2. In order to derive the two parameters mentioned above, we need to estimate several temporary parameter estimations using Eqs (4) to (9). These temporary estimators included the estimation of $\hat{z}_{ijk}$. μ_{zij}, μ_{zi}, σ_{zi}^2, and σ_μ^2. The detailed derivation can be found in the work of Bhaumik and Gibbons (2005).

$$\hat{z}_{ijk} = \frac{y_{ijk} - \alpha_i}{x_j}, \quad x_1 = 20\ ug/mL\ ;\ x_2 = 100\ ug/mL \tag{4}$$

$$\mu_{zij} = \frac{\sum_{k=1}^{5} \hat{z}_{ijk}}{5} \tag{5}$$

$$\mu_{zi} = \frac{\sum_{j=1}^{2} \mu_{zij}}{2} \tag{6}$$

$$\sigma_{zi}^{2} = \frac{\left(\sum_{j=1}^{2} \sum_{k=1}^{5} \left(\hat{z}_{ijk} - u_{zij} \right)^2 / (5-1) \right)}{2} \tag{7}$$

$$\sigma_{\mu}^{2} = \frac{\left(\sum_{j=1}^{2} \sigma_e^2 / x_j^2 \right)}{2} \tag{8}$$

After the calculation using Eqs. (4)-(9), we can obtain the estimation of β and σ_{η}^{2}.

$$\beta_i = \sqrt{\frac{\mu_{zi}^{4}}{\left(\sigma_{zi}^{2} + \mu_{zi}^{2} - \sigma_{\mu}^{2} \right)}} \tag{9}$$

$$\sigma_{\eta}^{2} = \frac{2\sum_{i=1}^{5} \ln\left(\mu_{zi} / \beta_i \right)}{5} \tag{10}$$

(C) Point Estimate of Unknown True Concentration X

If we submit the sample to new q' independent laboratories, and $Y_1, Y_2, \cdots, Y_{q'}$ are the corresponding measurements, we could easily derive the estimate of the unknown true concentration X from our model defined in Eq. (1).

We use the dataset in Table 1 to demonstrate the true concentration estimation procedure. After finishing the parameter estimation by using the whole dataset from the five laboratories, we choose

first three laboratories $q'=3$ to calculate the asymptotically unbiased estimator X with its variation in Equation (11) to (12).

$$X_i=\frac{Y_i-\alpha_i}{\beta_i\gamma} \text{ and } X=\sum_{i=1}^{3}\frac{X_i}{3}, \text{ where } \gamma=E\left(e^{\eta}\right)=e^{\frac{\sigma_\eta^2}{2}} \tag{11}$$

$$\mathrm{Var}\left(X\right)=\frac{\sum_{i=1}^{3}\frac{\sigma_e^2\left(1+1/5\right)}{\beta_i\hat{r}}}{3^2}+\frac{X^2\left(\hat{r}^2-1\right)}{3} \tag{12}$$

(D) Confidence Interval of Unknown True Concentration X

- The estimation of Confidence Interval of Low Concentration X_0

According to the characteristic of low concentration analyte, the measurement is close to the normal distribution. We define $\bar{Y}_0$ as the average value of the first three laboratories observation. Also, let σ_α^2 represent the variability of α_i among the five laboratories. Hence, we could obtain the $100\%\left(1-\alpha\right)$ confidence interval of X_0, as shown in Eq. (13).

$$\left(\max\left(0,\bar{Y}_0-z_{\alpha/2}\sqrt{\frac{\sigma_e^2+\sigma_\alpha^2}{3}}\right),\ \bar{Y}_0+z_{\alpha/2}\sqrt{\frac{\sigma_e^2+\sigma_\alpha^2}{3}}\right) \text{ , where } \bar{Y}_0=\sum_{i=1}^{3}\frac{Y_{i0}}{3} \tag{13}$$

- The estimation of Confidence Interval of High Concentration X

The measurement of high concentration analyte needs to follow the lognormal distribution. First, we define $V_i=\ln\left(\frac{Y_i-\alpha_i}{\beta_i X}\right)$ and obtain $Z_i\left(X\right)=\frac{V_i-E\left(V_i\right)}{Var\left(V_i\right)}=\frac{\ln\left(Y_i-\alpha\right)-\ln\left(\beta_i X\right)}{\sqrt{c_{3i}}}\sim N\left(0,1\right)$. After that, we can calculate the $100\left(1-\alpha\right)\%$ confidence interval of X for each specific laboratory. The

formula is shown in Eq. (14). Based on this concept, we further define $Z(X)$ to obtain the $100(1-\alpha)\%$ confidence interval of X, as defined in Eq. (15)

$$\left(\frac{1}{e^{\ln(\beta_i)-\ln(Y_i-\alpha_i)+z_{\alpha/2}\cdot c_{3i}}}, \quad \frac{1}{e^{\ln(\beta_i)-\ln(Y_i-\alpha_i)-z_{\alpha/2}\cdot c_{3i}}} \right) \tag{14}$$

$$R(X)=\left\{ X: \quad -Z_{\alpha/2} \ \le \ Z(X)=\frac{\sum_{i=3}^{3} Z_i(X)}{\sqrt{3}} \sim N(0,1) \ \le \ Z_{\alpha/2} \right\} \tag{15}$$

3. Data Analysis

(1) Exploratory Data Analysis and Model Selection

We first examine whether the environmental analyte has the following characteristics: (i) low concentration dataset (0 μg/L) is normal distribution, and (ii) high concentration dataset (20μg/L and 100μg/L) is lognormal distribution. This is performed by the Shapiro-Wilk normality test and the QQ-plot of standard residuals.

Two points are made from the results in Fig. 1. The first is that for the low-concentration data, it does not follow the normal distribution. This may be caused by the extreme values in the dataset, especially for the data collected by the third laboratory. The second point is that for the high-concentration data with the highest intensity (100μg/L), it agrees well with the lognormal distribution when applying data transformation in a log scale. However, for the other group of high-concentration data with mild intensity (20μg/L), it only shows an acceptable match with the lognormal distribution. A better agreement is reached if we remove the outliers.

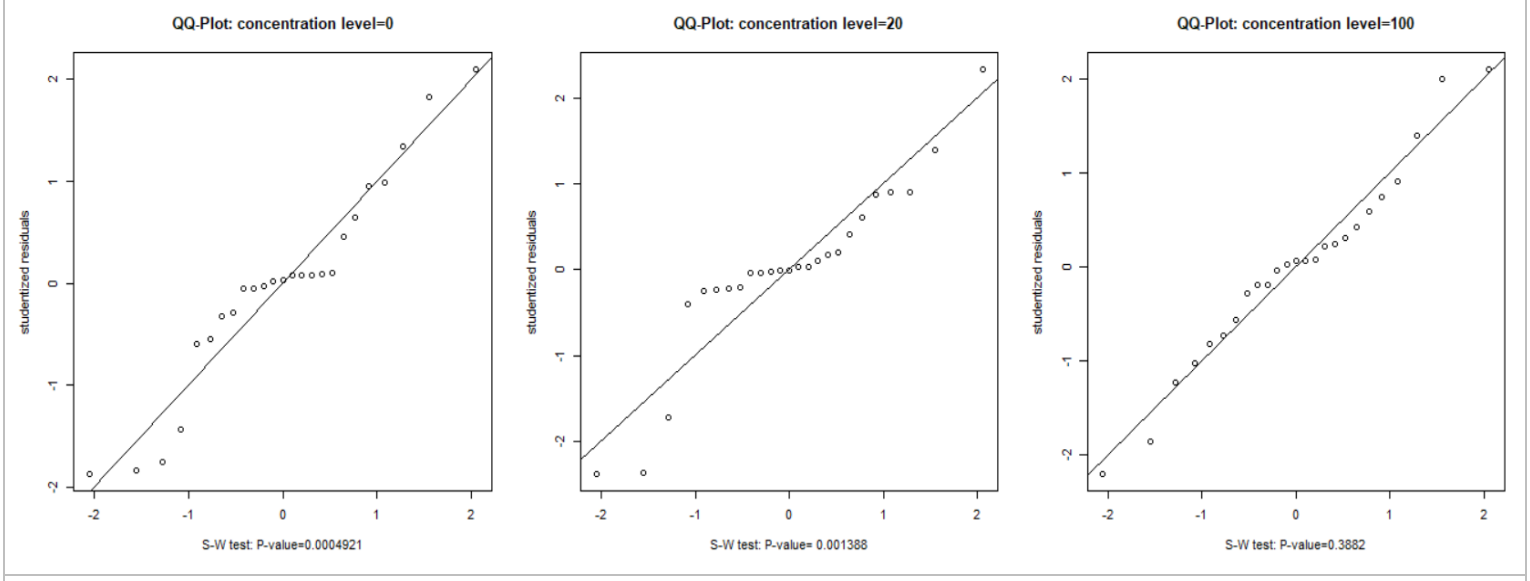

Figure 1. The normality performance among concentration level

Now we perform exploratory data analysis to decide the appropriate model for estimation. Fig. 2 indicates that a higher level of concentration exhibits a more significant variation of measurement. The variances are 14.1135, 15.0001, and 57.0405, respectively for the concentration levels being 0, 20, and 100μg/L. For the sake of fitting this type of dataset, we have to use the nonlinear model, $y_{ijk} = \alpha_i + \beta_i x_j e^{\eta_{ijk}} + e_{ijk}$ rather than the traditional linear model (Bhaumik and Gibbons (2005)).

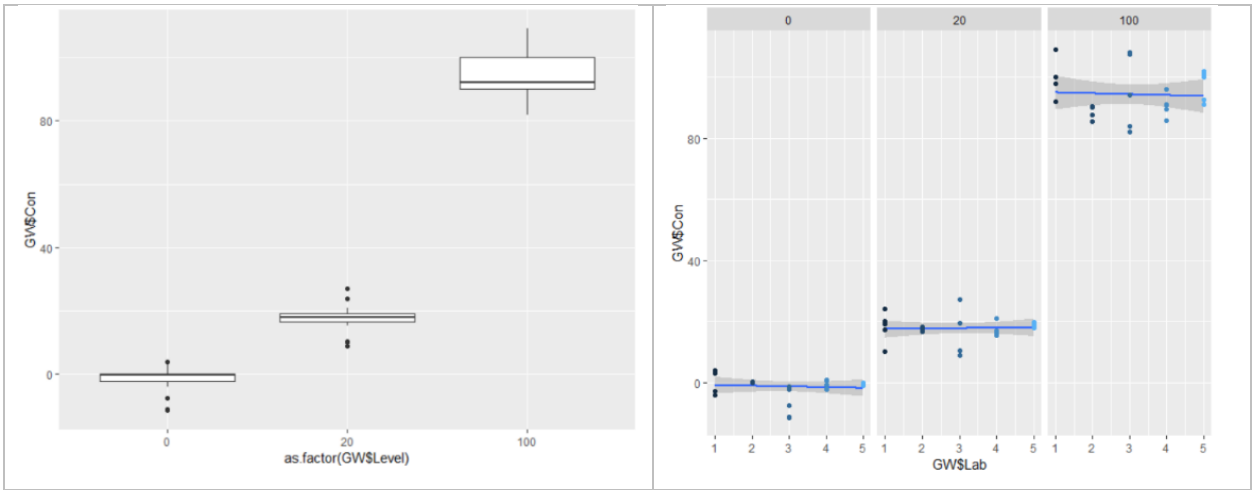

Figure 2. The variances are proportion to concentration level

(2) Parameter Estimation & True Concentration Level – For Cadmium Dataset

After choosing the type of the model, the next step is to estimate the model parameter. Here we use the methods of moments as suggested by Bhaumik and Gibbons (2005). The estimation process is performed using the R programming of the version of 3.5.1. Table 2 summarizes the result of the parameter estimation.

Table 2. Parameter Estimation

Lab	α	$Se(\alpha)$	β	$Se(\beta)$
1	0.6200		0.9187	
2	0.0700		0.8829	
3	-6.6800	3.1138	1.0735	0.0765
4	0.6308		0.9018	
5	-0.1924		0.9692	
σ_e^2	7.8955			
σ_η^2	0.0110			

Consequently, our models are,

$$\text{Laboratory 1: } Y_1 = \alpha_1 + \beta_1 Xe^{\eta} + \hat{e} = 0.6200 + 0.9187 Xe^{\eta} + \hat{e}$$

$$\text{Laboratory 2: } Y_2 = \alpha_2 + \beta_2 Xe^{\eta} + \hat{e} = 0.0700 + 0.8829 Xe^{\eta} + \hat{e}$$

$$\text{Laboratory 3: } Y_3 = \alpha_3 + \beta_3 Xe^{\eta} + \hat{e} = -6.6800 + 1.0735 Xe^{\eta} + \hat{e}$$

$$\text{Laboratory 4: } Y_4 = \alpha_4 + \beta_4 Xe^{\eta} + \hat{e} = 0.6308 + 0.9018 Xe^{\widehat{\eta} + \hat{e}}$$

$$\text{Laboratory 5: } Y_4 = \alpha_5 + \beta_5 Xe^{\eta} + \hat{e} = -0.1924 + 0.9692 Xe^{\widehat{\eta} + \hat{e}}$$

$$\hat{e} \sim Normal(0, \sigma_e^2) = Normal(0, 7.8955)$$

$$\widehat{\eta} \sim Normal(0, \sigma_\eta^2) = Normal(0, 0.110)$$

Now we can use the models to estimate the unknown true concentration X. Assume the new measurements are collected in the q' independent laboratories. From Eqs. (11) and (15), we calculate the point estimation, variance, confidence interval of X, and simulated confidence level (SCL).

There are two approaches to estimate X. The first approach follows the work of Bhaumik and Gibbons (2005). We use the first three laboratories (q'=3) and the first replicate sample from the laboratories (q=5). As shown in Table 3, the point estimate of the concentration level X=0 is -1.5773 with 95% confidence interval (0.0000, 1.1702). For X=20, the point estimate is 20.4786 with 95% confidence interval (15.4935, 23.1297). Finally, for X=100, the point estimate is 102.1374 with the 95% confidence interval (90.7669, 116.1490).

A potential drawback of the first approach is that the result may be biased if the first three laboratories have any extreme value, for instance, most of the data in the third laboratory present the extreme condition. To avoid this problem, we propose the second approach in this report. We randomly choose the three laboratories among the total of the five laboratories and then randomly assign one replicated measurement. Table 3 shows the results respectively after 10,000 draws. It indicates that with 10,000 draws of random sampling, the mean point estimate of true concentration is closer to the population estimation as opposed to the first approach.

In both approaches, the variances are proportional to the concentration level. Good performances are found since the simulated confidence levels (SCL) are close to 95%.

Table 3. True Concentration Estimation (q'=3)

True Concentration	Reproduce from Bhaumik (2015) Lab=1,2,3; Rep=1st				Data Set Random Sampling (S=10,000) Lab=Random; Rep=Random			
	X	$Var(X)$	CI	SCL	X	$Var(X)$	CI	SCL
0 μg/L	-1.5773	3.4728	0.0000, 1.1702	0.9375	0.0006	3.5716	0.0000, 3.3060	0.9706
20 μg/L	20.4786	4.9507	15.4935, 23.1297	0.9370	19.8719	4.9954	16.1382, 24.0922	0.9481
100 μg/L	102.1374	40.4201	90.7669, 116.1490	0.8921	100.4810	40.4662	89.1430, 114.0710	0.9057

(3) Parameter Estimation & True Concentration Level – Simulation for Cadmium Dataset

The final step is to perform the simulation using the previous result of parameter estimation in Table 2. The parameters are updated after executing a simulation. The updated parameters will be eventually converged to the original ones if the initial evaluation of parameters is correct, and vice versa. In this report, we repeat the simulation respectively with one and 10,000 times. Table 4 shows the results. As expected, we observe a decent agreement. When we repeat the simulation with 10,000 times, the mean of the parameter estimation well converges to the original ones (Table 2).

Table 4. Parameter Estimation from Simulated Data

Simulated Dataset Estimation (Simulation one time)					Mean and SD from Simulation (Simulation 10,000 times)				
Lab	α	$Se(\alpha)$	β	$Se(\beta)$	Lab	α	$Se(\alpha)$	β	$Se(\beta)$
1	1.0261		0.8954		1	0.6110	1.2753	0.9188	0.0591
2	1.4990		0.8180		2	0.0538	1.2653	0.8832	0.0580
3	-5.6267	2.9473	1.0911	0.1006	3	-6.6961	1.2522	1.0736	0.0617
4	0.6455		0.9075		4	-0.6305	1.2457	0.9019	0.0581
5	0.4332		0.9084		5	-0.1809	1.2392	0.9683	0.0591
σ_e^2	7.1447				σ_e^2	7.7412	2.3769		
σ_η^2	0.0083				σ_η^2	0.0115	0.0062		

Likewise, after obtaining the new parameters, we re-calculate the estimation of the true concentration level X, as shown in Table 5. The results show a good agreement again. For a simulation repeated 10,000 times, the values of the updated X gradually converge into the original X (Table 3). From the simulation results, the expected characteristics of the nonlinear model is confirmed, namely, the variances are found to proportional to the concentration level.

Table 5. True Concentration Estimation (q'=3) form the Simulated Dataset

True Concentration	Simulated Dataset (Simulation one time) Lab=Random; Rep=Random			Simulated Dataset (Simulation 10,000 times) Lab=Random; Rep=Random		
	$\hat{X}$	$Var(\hat{X})$	CI	$\hat{X}$	$Var(\hat{X})$	CI
0 μg/L	-0.9271	3.4702	0.0000, 6.7047	-0.0494	3.4850	0.1082, 2.1038
20 μg/L	19.4762	4.2321	15.7908, 23.0489	19.8691	5.0883	16.2574, 24.2951
100 μg/L	96.6616	31.5633	86.4945, 107.3899	100.3703	43.4056	88.9287, 113.8584

4. Conclusion

In this report, we construct a nonlinear model to fit the Cadmium dataset following the work of Bhaumik and Gibbons (2005). For the parameter estimation, we use the method of moments because the sample moments can provide good estimates without the burden of time penalty. To reduce the measurement uncertainty, we use all the observations among five laboratories to establish the parameter estimation procedure. Through taking interlaboratory variation will make the perdition of unknown true concentration level more pragmatic. Finally, the asymptotically property is also verified by comparing with our simulation result.

5. Reference

[1] Bhaumik, D.K., and Gibbons, R.D., Confidence Regions for Ranom-Effects Calibration Curves With Heteroscedastic Errors," *Technometrics*, 40, 223-230

[2] Gibbons, R.D. (2001), *Statistical Methods for Detection and Quantification of Environmental Contamination*, New York: Wiley.

[3] Gibbons, R.D., Bhaumik, D.K., Aryal, S.,(2009) *Statistical Methods for Groundwater Monitoring*, New York: Wiley.

[4] Rocke. D, M., and Lorenzato, S. (1995), "Two-Component Model for Measurement Error in Analytical Chemistry," *Technometrics*, 37, 176-184