Technical Basis for the Lowest Concentration Minimum Reporting Level (LCMRL) Calculator

Office of Water (MLK 140) EPA 815-R-11-001 December 2010 http://water.epa.gov/drink

Table of Contents

1.0 Introduction

1.1 Programmatic Background

The Safe Drinking Water Act Amendments of 1996 require EPA to establish criteria for a monitoring program and to publish a list of not more than 30 unregulated contaminants for which public water systems (PWS) are to monitor. The monitoring program will provide a national assessment of the occurrence of these contaminants in public drinking water that will be used to help decide which contaminants may or may not require regulation in the future. In 1999, EPA revised the approach for unregulated contaminant monitoring in the Unregulated Contaminant Monitoring Regulation (UCMR) (64 FR 50556, USEPA, 1999) and subsequent revisions.

PWSs will be required to monitor for a variety of contaminants under the UCMR. A Minimum Reporting Level (MRL) will be assigned to each contaminant. Under the UCMR, laboratories will be required to report all occurrences of listed contaminants at concentrations that are equal to or greater than the established MRL.

MRLs represent an estimate of the lowest concentration of a compound that can be quantitatively measured by members of a group of experienced drinking water laboratories. It is based on a Measurement Quality Objective (MQO) of 50%-150% recovery of spiked concentrations. Informally, at or above the MRL, competent drinking water laboratories should be expected to obtain 50%-150% recovery or better.

In both the Information Collection Rule (ICR) and the UCMR, the EPA Office of Ground Water and Drinking Water (OGWDW) specified MRLs and an accuracy requirement for recovery at the MRL. The MRL was introduced with the new analytes and new methods for implementing the new UCMR.

The Method Detection Limit procedure (MDL; see 40 CFR part 136, Appendix B) was developed by EPA (Glaser et al., 1981) to establish a decision threshold for detection in low-level samples. It is based solely on the standard deviation of repeated measurement of low-level spikes. The MRL differs from the MDL by considering not only the standard deviation of low concentration analyses (precision) but also the accuracy of the measurements as they impact achievement of the MQO for spike recovery.

The practical quantitation limit (PQL) was established by EPA's drinking water program. The PQL is defined as "the lowest concentration of an analyte that can be reliably measured within specified limits of precision and accuracy during routine laboratory operation conditions" (52 FR 25690, USEPA, 1987). The PQL is problematic as a practical matter since at least three different methods have been used to determine PQLs:

1. analysis of performance evaluation (PE) sample data,
2. a risk-based multiple of the MDL, and
3. adoption of Contract Laboratory Program Contract Required Quantitation Limits (CRQLs) based on lowest nonzero standard in a linear calibration.

The MRL may be useful as an alternative to the PQL for setting future regulatory limits.

A necessary element of determining the MRL is a procedure to estimate the lowest level at which a single laboratory can meet the specified MQOs. EPA has developed a statistical approach for determination of single-laboratory Lowest Concentration MRLs (LCMRLs). This approach uses variance function modeling, regression and modeling instrument response distributions.

The MRL will be determined using a Bayesian bootstrap (BB) (Rubin, 1981) of the LCMRL estimator using the LCMRL study data from each of several experienced drinking water laboratories. The BB replicates that were generated from each laboratory's data, serve to estimate the distribution of estimated LCMRL values that each laboratory might generate on repeated performance of the LCMRL study. The distribution of pooled BB replicates, generated from the LCMRL study data from a sample of experienced drinking water laboratories, approximates the distribution of estimated LCMRL values which might be generated from the *population* of experienced drinking water laboratories.

Reasonable values for programmatic required MRLs can be determined based on these simulated distributions. Drinking water laboratories will confirm that they are capable of meeting a required MRL during their Initial Demonstration of Capability (IDC) for an analytical method. A technical description of the MRL procedure is given in Appendix A.

1.2 Scientific Background

The *Lowest Concentration Minimum Reporting Level* (LCMRL) is defined as the lowest *spiking* concentration such that the probability of spike recovery in the 50% to 150% range is at least 99%. The algorithms and procedure presented in this report provide a means of reliably estimating the LCMRL.

The LCMRL calculator constructs conditional (on spiking concentration) mean and variance models for analytical measurements. For analytical methods which cannot give negative 'raw' measurements or negative results (except possibly at low level due to a negative intercept in the calibration function), the gamma distribution is used as the distribution of measurements. Reported negative measurements due to a calibration artifact are set to zero. Typically, these analytical methods are chromatographic methods in which the response is integrated valley to valley.

Other classes of methods can obtain 'raw' measurements at very low concentration that are negative. These methods may be spectroscopic or radiochemical methods and often involve subtraction of 'dark current,' reference cell readings or background. In these cases the normal distribution is used as the distribution of measurements.

The conditional mean and variance models then specify the parameters for the conditional measurement distribution as a function of spiking level. It is this model of the distribution of repeated measurements at a given spiking level that is used to estimate the LCMRL.

The *critical level*, L_c is defined as an upper percentile of the distribution of repeated measurements at zero true concentration (see Section 10.1, and Currie, 1968). The modified Hubaux-Vos Detection Limit (mHV-DL) is defined as the spiking concentration that has a 95% probability of exceeding the 95% L_c. The conditional distributional model constructed for determining the LCMRL is also used to determine the mHV-DL.

1.3 Improvements over Previous LCMRL Methodology

The currently proposed LCMRL methodology offers several improvements over the previous LCMRL (USEPA, 2004).

The first improvement is the fundamental use of a probability-based definition for the LCMRL in deriving the estimation algorithms. As described above, the LCMRL is defined as the lowest *spiking* concentration such that the probability of spike recovery in the 50% to 150% range is at least 99%. The previous LCMRL calculator used a graphical technique which did not completely agree with the probability- or MQO-based definition of the LCMRL. The probability-based definition leads to an approach based on characterizing the distribution of the analytical response as a function of the spiking level.

The second improvement is that the piecewise linear replicate variance function in the first LCMRL calculator has been replaced by a smooth parametric function. The piecewise linear model was based on linear interpolation of pointwise sample variances. The proposed methodology replaces the ordinary sample variances with robust estimates of variance based on modern statistical methods (M-estimation). This enables the potential impact of outliers to be effectively eliminated without requiring outlier testing and removal.

The third improvement is that the robust replicate variance estimates are fit to a model based on an additive and multiplicative error structure for trace level analytical chemistry measurements.

The fourth improvement is that the average response model (regression of measured onto spiked concentrations) is fit using iteratively reweighted least squares (IRLS) in which the weights are determined not only by the reciprocal of the robust variances but also by M-estimation weights determined as part of the robust variance estimation process. This again enables the potential impact of outliers to be effectively eliminated without requiring outlier testing and removal.

The fifth improvement is in considering lack of fit for the average response model as a component of prediction variance. This is done by computing robust estimates of mean squared error (MSE) and fitting a parametric model for conditional MSE, similar to what is done for replicate variance. The prediction variance function is constructed using the pointwise maximum of the replicate variance and conditional MSE functions. This is multiplied by a function that includes the additional uncertainty inherent in the estimation of the average response model. This is a conservative approach based on the author's perception that the greater danger arises from underestimating the prediction variance.

The sixth improvement is that the gamma distribution is used as a response model for analytical methods which do not give negative results at low level. The normal distribution is used as the

response model for analytical methods which can give negative results at low level. The degrees of freedom in the normal case are estimated from the approximate degrees of freedom for the variance predictions given by the replicate variance model.

The seventh improvement is using the conditional response distribution estimated to determine the LCMRL to estimate L_c and the detection limit using a modification of the Hubaux-Vos approach.

2.0 General Notation

Below are definitions of key variable notations used throughout this paper. Additional notations included in this document are described as they are introduced.

m is the number of spiking levels.

n is the total number of observations.

n_i is the number of replicate observations at the i^{th} spiking level.

x_i is the spiking concentration for the i^{th} spiking level.

y_{ij} is the measured concentration for the j^{th} measurement at the i^{th} spiking level.

Y_{ij} is the j^{th} measurement at the i^{th} spiking level represented as a random variable.

w_{ij} is the weight for the j^{th} measurement at the i^{th} spiking level.

$\underline{y}$ is the nx1 vector of measured concentrations.

p is the number of model parameters.

$\underline{\beta}$ is a px1 vector of regression coefficients (model parameters).

$E(.)$ is the expected value function for a random variable.

var(.) is the variance function for a random variable.

μ_i is the expected value of Y_{ij}.

σ_i^2 is the variance of Y_{ij}.

Z is a design matrix.

$\underline{Z}_i$ is a design vector (row of the design matrix) for the measurements at the i^{th} spiking level.

CV is coefficient of variation (also known as relative standard deviation).

L_c is the critical level.

γ_C is the coverage probability for the critical level, L_c.

γ_D is the coverage probability for the mHV-DL.

γ_Q is the coverage probability for the quality specification interval (50%-150% recovery) at the LCMRL.

3.0 LCMRL Study Design

Because the LCMRL methodology is a regression-based methodology, LCMRL study designs involve replicate spiking at multiple levels. The relevant design parameters are:

- the number of spiking levels,
- the spacing of spiking levels,
- the number of replicates at each spiking level (not necessarily equal), and
- inclusion of method blanks.

The relevant issues associated with the design of detection and quantitation studies include the impact of the study design on:

- identification of nonlinearity and the correct model form for the average response function,
- parameter estimation in the replicate variance and conditional MSE model, and
- parameter estimation in the average response function model.

Because the LCMRL methodology is complex and multistage, definitive recommendations for study designs are not available at present. This is an issue for further development.

Hubaux and Vos (1970) study a variety of standards spacings, which they refer to as repartitions of standards. They study linear spacing, parabolic spacing, two point spacings (that is, two spiking levels) and three point spacings. They recommend that the ratio of the maximum to minimum (nonzero) spiking level be about 10. They also recommend having all the replicates (save one) at the lowest and highest spiking levels. This is their three point design. This report rejects that particular recommendation for the following reasons.

Both the average response and replicate variance models are regression models. In design for regression, having more levels of the predictor (the spiking level in this case) is much more important than having more replication at each level. However, for the LCMRL, estimation of the replicate variance function is also vital. Therefore, some replication is also essential. Under Gaussian sampling, the coefficient of variation of the ordinary sample variance is 1 for three replicates, $\sqrt{2/3} \approx 0.82$ for four replicates, $\sqrt{2/5} \approx 0.63$ for six replicates, $\sqrt{2/7} \approx 0.53$ for eight replicates and $\sqrt{2/9} \approx 0.47$ for ten replicates.

Nonlinearity of the response function at low concentration is a well known phenomenon. It is very important to characterize nonlinearity at low level in order to accurately estimate the LCMRL and DL because the mean response function at low level is a key component of the LCMRL/mHV-DL methodology. For this reason, we allow a polynomial model, up to a cubic, for the mean response. In order for the coefficients of a cubic polynomial to be identifiable (estimable), there must be at least four spiking levels.

It is also very important to estimate the variance function as accurately at low level as possible. For these reasons, this report recommends the parabolic design for spiking levels with as many levels as feasible and at least four replications per spiking level.

The spiking levels may be computed as follows

$$x_i = x_1 + (x_m - x_1)\left(\frac{i-1}{m-1}\right)^2, \quad \text{for } i = 1,\ldots,m, \qquad (1)$$

where x_1 and x_m are the designed minimum and maximum spiking levels.

The proposed LCMRL calculator recommends seven spiking levels with four replicates per level for a total of 28 analyses. Fitting a cubic polynomial in the mean response model requires a **minimum** of four spiking levels. The minimum values that the LCMRL calculator will accept are four spiking levels and three replicates per level. Given that the size of the LCMRL/DL study must be limited by cost and resource availability, a tradeoff must be made between estimating replicate variance and MSE at individual spiking levels and estimating the regression models for mean response, replicate variance and MSE. The recommended numbers of spiking levels and replicates are a reasonable compromise.

4.0 Data Conditioning

Because the LCMRL methodology is a complex, multistage, model-based design, care must be taken in screening the input data for use in the model development. The model estimation procedure is sensitive to departures from a monotonic variance model. Having non-zero spiking levels with non-zero replicate variances is problematic for modeling. Therefore, data values that would contribute to invalidating these assumptions are dropped from the estimation procedure. Data are dropped from the estimation procedure for non-zero spiking levels where 50% or greater of replicate responses are zero. If less than 50% of the replicate responses of a non-zero spiking levels are zero, then the zero responses are replaced by the minimum non-zero replicate response within the spiking level or within a lower spiking level.

After data conditioning the procedure determines if the requirement of a minimum of four spiking levels and three replicates is achieved before estimation of the LCMRL.

5.0 Estimation of Replicate Variance

5.1 Introduction

Because of the potential for outliers, replicate variance at each spiking level is estimated using weighted variances using weights computed by M-estimation. As a byproduct of this process, robust weights are available for estimating the conditional mean (regression) model. Since at each spiking level the number of results may vary from 3 to perhaps 10 or more, this is a very challenging undertaking.

The objective is not to 'identify' outliers as such but merely to down-weight them sufficiently that they do not overly influence the means model, replicate variance model or conditional MSE model, without overly compromising efficiency. In contrast, testing and removal of outliers constitutes a data weighting scheme where the only possible weights are 0 or $1/n$, where n is the number of retained observations. This is very inflexible compared to weighting based on M-

estimation. Not only is testing and removal of outliers awkward procedurally, it causes a loss of statistical efficiency.

A modified Hodges-Lehmann estimator is used for the initial resistant location estimate. The scaled mean absolute deviation from the Hodges-Lehmann estimator is used as the resistant scale estimator. Using this initial location-scale estimation pair, the Huber M-estimate of location and its associated weights are computed. A robust scale estimator is computed based on the weighted mean absolute deviation from the Huber location estimator and using the Huber weights. The Huber location-scale estimates are then used as an initial location and robust scale estimates for Tukey's biweight procedure.

Finally, the Tukey's biweight weights are used to compute location and variance estimators at each spiking level. The weights are also later used in computing the conditional mean response (regression) model.

5.2 Modified Hodges-Lehmann Location Estimator

The Hodges-Lehmann location estimator (Randles and Wolfe, 1991) is the median of the set of means of all pairs of observations. The modified Hodges-Lehmann location estimator (modified for this application) at the i^{th} spiking concentration combines the median of the measurement data with the set of means of pairs of data and takes the median of the resulting set of numbers. For a set of n_i observations $\left\{y_{i1},\ldots,y_{in_i}\right\}$ for the i^{th} spiking level, this is given by

$$m_{HL,i} = \text{median}\left\{\bigcup_{j<k}\frac{y_{ij}+y_{ik}}{2}, \text{median}\left\{y_{i1},\ldots,y_{in_i}\right\}\right\}. \quad (2)$$

5.3 Mean Absolute Deviation Scale Estimator

This estimator is based on the long known (Kelley, 1921) Absolute Deviation (AD) scale estimator. It uses the mean absolute deviation from the Hodges-Lehmann estimator (instead of from the sample mean). It is computed as

$$AD_i = 1.4826 \cdot \tfrac{1}{n}\sum_{j=1}^{n_i}\left|y_{ij} - m_{HL,i}\right|. \quad (3)$$

The factor 1.486 is a normalization constant, used so that the expected value of the AD in normal samples equals the population standard deviation.

5.4 Huber Location Estimator and Weights

The Huber estimator (Hoaglin et al., 1983) is calculated iteratively as a weighted estimator (*w*-estimator) using a relative convergence criterion of 10^{-6}. The process is started with a resistant initial location estimate $T_{H,i}^{(0)}$ given by the modified Hodges-Lehmann estimator and a resistant auxiliary scale estimate given by the modified AD estimator, AD_i. This is a departure from the

usual practice, in which the median is used as the initial location estimate and the median absolute deviation (MAD) scale estimator is used as the auxiliary scale estimator for the Huber location estimator. The modification is prompted by the very small sizes of the data sets in this application.

The Huber weights and estimator are determined by a tuning constant c_H according to the equations below. The value of the parameter c_H is taken to be 1, which is a standard value often used in practice. The index k tracks the iteration number. For a set of n_i observations $\left\{y_{i1},\ldots,y_{in_i}\right\}$ at the i^{th} spiking level the iteration proceeds as follows

$$
\begin{aligned}
u_{ij}^{(k)} &= \left|\frac{y_{ij}-T_{H,i}^{(k-1)}}{AD_i}\right|,\ j=1,\ldots,n_i \\
w_{ij}^{H,k} &= \begin{cases} 1, & \text{if } u_{ij}^{(k)} \le c_H \\ c_H \big/ u_{ij}^{(k)}, & \text{if } u_{ij}^{(k)} > c_H \end{cases}, \quad j=1,\ldots,n_i \\
T_{H,i}^{(k)} &= \left(\sum\nolimits_{j=1}^{n_i} w_{ij}^{H,k} y_{ij}\right) \Big/ \sum\nolimits_{j=1}^{n_i} w_{ij}^{H,k}
\end{aligned}
\tag{4}
$$

until the estimate $T_{H,i}^{(k)}$ converges, at which time the weights are assumed to have converged based on a relative convergence criteria of 10^{-4}. The Huber M-estimate of location is denoted by $T_{H,i}$ and the associated weights by w_{ij}^{H}.

The Huber scale estimator, $s_{H,i}^2$ (equation 5), is computed as the weighted variance around the Huber location estimate using the Huber weights. This formula is derived from the variance of a weighted mean. It also follows from an approximation to the A-estimators of scale (Lax, 1985, Section 4.6.2) based on the asymptotic variance of M-estimators (Hoaglin et al., 1983). This completes the circle, since M-estimators can always be represented as weighted means (W-estimators).

$$
\begin{aligned}
n_{w,i} &= n_i\left[1-\sum_{j=1}^{n_i}\left(w_{ij}^{H}\right)^2\right] \\
s_{H,i}^2 &= \frac{n_i}{n_{w,i}}\sum\nolimits_{j=1}^{n_i} w_{ij}^{H}\left(y_{ij}-T_{H,i}\right)^2
\end{aligned}
\tag{5}
$$

5.5 Tukey's Biweight Location Estimator and Weights

The Tukey's biweight M-estimator (Hoaglin et al., 1983; Horn, 1988) is calculated iteratively as a weighted estimator using a relative convergence criterion of 10^{-4}.

The biweight is more robust statistically in certain respects than the Huber estimator. It is very efficient and can down-weight extreme observations to zero. On the other hand, it is not as robust computationally as the Huber estimator. Without a good auxiliary scale estimate and a

good starting location estimate, it can misbehave. To address this issue the process is started with a robust initial location estimate $T_{BW}^{(0)}$ given by the Huber estimator and a robust auxiliary scale estimate given by Huber weighted AD scale estimator, $s_{H,i}$. This is a departure from the usual practice, in which MAD is used as the auxiliary scale estimator. It is prompted by the very small sizes of the data sets in this application.

The Tukey weights are determined by a tuning constant c_{BW} according to the equations below. The value of the tuning parameter c_{BW} is taken to be 9. The index k tracks the iteration number. For a set of n_i observations $\left\{y_{i1},\ldots,y_{in_i}\right\}$ at the i^{th} spiking level the iteration proceeds as follows

$$
\begin{aligned}
u_{ij}^{(k)} &= \tfrac{y_{ij}-T_{BW,i}^{(k-1)}}{c_{BW}s_{H,i}},\ j=1,\ldots,n_i \\
w_{ij}^{BW,k} &= \begin{cases} \left(1-\left(u_{ij}^{(k)}\right)^2\right)^2, & \text{if } u_{ij}^{(k)}<1 \\ 0, & \text{if } u_{ij}^{(k)}\geq 1 \end{cases},\ j=1,\ldots,n_i \\
T_{BW,i}^{(k)} &= \left(\textstyle\sum_{i=1}^{n_i} w_{ij}^{BW,k} y_{ij}\right) \Big/ \textstyle\sum_{i=1}^{n_i} w_{ij}^{BW,k}
\end{aligned}
\qquad (6)
$$

until the estimate $T_{BW,i}^{(k)}$ converges, at which time the weights are assumed to have converged. The location estimate is denoted by $T_{BW,i}$ and the weights by w_{ij}^{BW}. Note that the definitions of $u_{ij}^{(k)}$ are similar but different in the two cases (Huber and biweight).

5.6 Final Location-Scale Estimator and Weights

The final robust weights are the Tukey biweights. The robust location and variance estimates for the i^{th} spiking level are computed as a weighted mean, $m_{w,i}$, and weighted variance, $s_{w,i}^2$, of the data as indicated below where nw_i is the degrees of freedom associated with the weighted observations.

$$
\begin{aligned}
n_{w,i} &= n_i\left[1-\textstyle\sum_{j=1}^{n_i}\left(w_{ij}^{BW}\right)^2\right] \\
m_{w,i} &= \left(\textstyle\sum_{j=1}^{n_i} w_{ij}^{BW} y_{ij}\right) \Big/ \textstyle\sum_{j=1}^{n} w_{ij}^{BW} \\
s_{w,i}^2 &= \frac{n_i}{n_{w,i}} \textstyle\sum_{j=1}^{n_i} w_{ij}^{BW}\left(y_{ij}-m_{w,i}\right)^2
\end{aligned}
\qquad (7)
$$

As noted above, this is the formula for a weighted variance, derived from the variance of a weighted mean. It is also an approximation to the A-estimators of scale based on the asymptotic variance of Tukey's biweight location estimator.

Characteristics of the Huber and biweight estimators are combined by using the Huber estimates as scale and starting location in computing the biweight. The reason for doing this is that for M-estimators with monotone ψ-functions, like the Huber, the equation defining the M-estimator is guaranteed to have a unique solution. For redescending estimators, like Tukey's biweight, uniqueness is not guaranteed. This problem is particularly worrisome given the small data sets being used (results at each spiking level). Since the computer code for LCMRL/mHV-DL estimation must run unsupervised, the Huber estimate is a safe starting point for the biweight iterations.

6.0 Replicate Variance Model

6.1 Derivation

The replicate variance model, which models replicate variance as a function of spiking concentration, is based in part on the paper by Rocke and Lorenzato (1995). This paper identifies sources of both additive and multiplicative random error in trace level analytical chemistry measurements. Corresponding to this conceptual model, the LCMRL procedure fits constant plus power models to replicate variance and conditional MSE models. The constant term estimates the additive sources of variance and the power model term estimates the multiplicative sources. The power model term accommodates error variance behavior in higher concentration ranges that varies from near constant variance to shot noise to constant coefficient of variation (CV).

The variance at extremely low concentrations may be distorted by processes such as thresholding, smoothing, rounding and truncation. Smoothing, which is often used in spectroscopic, chromatographic, and radiologic methods, reduces measurement variance, especially for low-level measurements. Rounding and truncation also reduce measurement variance.

Response thresholding, which is used in chromatographic methods to manage data volume, can either reduce or increase measurement variance depending on the situation. Thresholding effectively replaces noise (except for the extreme upper tail of the noise distribution) and some low level signal with zeros. Thresholding then functions as a Type II censoring mechanism with an unknown censoring point (unless the laboratory reports the instrument settings).

Instrument thresholds have both direct and indirect impacts on estimating detection and quantitation limits. The main direct impact is that it is not possible to estimate the standard deviation of measurements at zero. However, by definition, the standard deviation at zero is required to calculate L_c (Currie, 1968). The EPA MDL procedure (Glaser et al., 1981) was constructed to deal with this problem by providing a way to estimate a standard deviation at a low concentration, and included instructions for determination of a concentration as close to zero as is possible that will generate a measurement.

Theoretically, in the absence of such distortion, low-level measurement variance (even for method blanks) would always be positive. For this reason, method blank data are not used in estimating the replicate variance model.

The replicate variance model takes the form

$$\sigma^2(x) = a_v + b_v x^{c_v}, \quad (8)$$

where x is the spiking concentration.

The replicate variance model parameters are estimated using a constrained iterated downhill Simplex Method minimization routine (Nelder and Mead, 1965) with starting values for b and c found by first fitting the robust variances using log-scale simple regression. Because of the potential for outliers, the ordinary sample variance is not used as the response in the fitting. Instead a robust estimate of variance is computed based on M-estimation, as described above.

The starting values b_0 and c_0, for b and c, are first found by fitting a reduced variance model of the form

$$\ln[\sigma^2(x)] = \ln(b_v) + c_v \ln(x),$$

using ordinary regression using the robust variances $s^2_{w,i}$ at each spiking level x_i.

Following this, the full replicate variance model is fit using an iterated constrained downhill Simplex Method with starting values defined as

$$a_s = \max\left(\frac{\sum_{i=1}^{2} n_{w,i} s^2_{w,i}}{\sum_{i=1}^{2} n_{w,i}}, 1e-8\right)$$

$$b_s = b_0$$

$$c_s = \min\left(\max\left(0, c_0\right), 2\right)$$

and a constrained loss function defined at each iteration j of the optimization as

$$loss = \sum_{i=1}^{n} e_i$$

where

$$e_i = \begin{cases} 10^{12}, \text{ if } a_j < 0 \text{ or } b_j < 0 \text{ or } c_j < 0 \text{ or } c_j > 2 \\ n_{w,i} \dfrac{\left(s_i^2 - \sigma_i^2\right)^2}{\sigma_i^2}, \text{ otherwise} \end{cases}$$

$\sigma_k^2(x) = a_k + b_k x^{c_j}$, at the k^{th} iteration of the optimization.

The optimization continues for estimates of *a*, *b* and *c* until the loss function is less than 10^{-16}. Normal variances (that is, the statistic computed from data) have their sample variances proportional to their means. The robust variances used in the model behave similarly. It is this

that motivates the choice of the functional form of the loss function. The large loss incurred when parameter estimates are out of valid ranges forces the Simplex Method to constrain the parameter values to valid ranges.

7.0 Conditional Means Model

The conditional means model is estimated using a robust estimation procedure that implements a nested Iterative Reweighted Least Squares (IRLS) procedure. As part and as a consequence of the iterations, the conditional mean squared error model and the robust weights are also updated until convergence of the procedure. The IRLS procedure can handle non-standard regression assumptions such as non-homogenous variances and non-normal errors. This method is also resistant to outliers in the data set by using weights that are derived using Tukey's bisquare method (See Section 4.5).

The general conditional means model is

$$\mu(x_i) = \mathrm{E}(Y_{ij}) = \underline{\beta}^T \underline{Z}_i \qquad \underline{Z}_i^T = (1, x_i, x_i^2, x_i^3) \tag{9}$$

To solve this in the IRLS framework, it is necessary to solve

$$\sum_{i=1}^{n} w(u_i)(y_i - \hat{\beta} z_i) z_i = \sum_{i=1}^{n} w(u_i) e_i z_i,$$

where e_i is the residual at each point and $w(u_i)$ are the Tukey bisquare weights (estimated independently here for the means model). The Tukey weights, defined in Section 5.5, are in this case a function of the x_i, the predicted values as an estimate location, the conditional mean squared error as an estimate of scale and a tuning constant of 9. The conditional mean square error is used instead of the traditional variance to allow for the model lack of fit.

It is first necessary to have initial starting values for $\hat{\underline{\beta}}$ and the conditional mean square error (see Section 8.0). Using the initial $\hat{\underline{\beta}}$, say $\hat{\underline{\beta}}_0$, the initial predicated values, residuals and the conditional mean square error (see Section 8.0) are calculated. The initial $\hat{\underline{\beta}}_0$ is found by performing weighted least squares regression using the robust weight matrix W (see Section 5.6). The solution for $\hat{\beta}_0$ using weighted least squares regression is given below.

$$\underline{\hat{\beta}}_0 = \left(\mathbf{ZWZ}^T\right)^{-1}\mathbf{ZW}\underline{y}$$

$$\mathbf{W} = \text{diag}\left\{w_{1,1},\ldots,w_{m,n_m}\right\}$$

$$\underline{y} = \left(y_{1,1}, y_{1,2},\ldots, y_{1,n_1},\ldots, y_{m,1},\ldots, y_{m,n_m}\right)^T$$

$$\mathbf{Z} = \begin{pmatrix} \underline{Z}_1^T \\ \vdots \\ \underline{Z}_1^T \\ \vdots \\ \underline{Z}_m^T \\ \vdots \\ \underline{Z}_m^T \end{pmatrix} \begin{matrix} \left.\begin{matrix}\\ \\ \\ \end{matrix}\right\} n_1 \text{ rows} \\ \\ \left.\begin{matrix}\\ \\ \\ \end{matrix}\right\} n_m \text{ rows} \end{matrix}$$

The nested IRLS is implemented for estimating $\underline{\hat{\beta}}$ as follows. Parameters epsilon and maxIter are 10^{-6} and 100, respectively.

1) From the initial $\underline{\hat{\beta}}_0$ calculate the predicated values and residuals.
2) For the outer loop from the initial residuals compute the initial estimate of the conditional MSE function, mse_0.
3) For the inner loop compute the initial Tukey weights W_0 using the predicted values and conditional MSE function mse_0 and a tuning constant of 9.
4) Use weighted least squares to obtain the estimate $\underline{\hat{\beta}}_i$.
5) Use the parameter estimate $\underline{\hat{\beta}}_i$ to obtain new predicated values and residuals.
6) Go back to step 3 until the estimate $|\underline{\hat{\beta}}_i - \underline{\hat{\beta}}_{i-1}|$ converges pointwise such that the maximum pointwise difference is less than epsilon or the number of iterations exceeds maxIter. Set $\underline{\hat{\beta}} = \underline{\hat{\beta}}_i$ and go to step 7.
7) Go to step 2 and compute a new conditional mean square error function mse_j using $\underline{\hat{\beta}}$ and the residuals from step 5.
8) If the maximum pointwise difference of $|\underline{\hat{\beta}} - \underline{\hat{\beta}}_0|$ (over the outer loop) is greater than the convergence tolerance and the maximum number of outer iterations has not been exceeded, set $\underline{\hat{\beta}}_0 = \underline{\hat{\beta}}$ and $mse_0 = mse_j$ and go back to step 1. Otherwise, stop and return the current values $\underline{\hat{\beta}}$ and mse_j.

Final model selection (linear, quadratic or cubic) is accomplished using Mallow's C_p (Draper and Smith, 1981). Since the constant term in the model may be negative, it may be possible for mean concentrations near zero to have a negative mean estimate. Therefore, the mean response function at a spiking concentration, *x*, is defined as

$$\mu(x) = \begin{cases} 0, \text{ if } \underline{\hat{\beta}} z \le 0 \\ \underline{\hat{\beta}} z, \text{ if } \underline{\hat{\beta}} z > 0 \end{cases}$$

8.0 Conditional Mean Squared Error Model

8.1 Estimating Conditional Mean Squared Errors

Since the sampling distribution for repeated measurements at a spiking level is based on the conditional means (regression) model and variance about the conditional mean, the contribution to variance by lack of fit must be accounted for. To account for the effect of lack of fit for the regression model, a model is fit for conditional mean squared error (cMSE) as a function of spiking level.

First, robust estimates are made of MSE at each spiking level. This process is analogous to the estimation of replicate variance except that the raw residuals from the conditional means model at each spiking level are used as the data. The robust cMSE for the i^{th} spiking level, x_i, is calculated as

$$cMSE_i = m_{w,i}^2 + s_{w,i}^2 \qquad (10)$$

with the quantities calculated as described in Section 5.6 but using the regression residuals. Clearly, the cMSE should be larger than the replicate variance at each spiking level.

8.2 Fitting Conditional Mean Squared Errors Model

The conditional mean squared error model, which models mean squared error as a function of spiking concentration, is assumed to be in the class of constant + power models

$$\tau_i^{\,2} = a_e + b_e x_i^{c_e}, \qquad (11)$$

where $i = 1, 2, \ldots, m$ indexes the spiking levels, x_i.

The cMSE model is fit in exactly the same manner as the replicate variance model. Because of the model fitting process however, there is no guarantee that the expected value from the cMSE model is larger than the expected value of the replicate variance model at *every* spiking level.

9.0 Estimating the LCMRL

9.1 Introduction

To construct an estimated sampling distribution for repeated measurements at an arbitrary spiking concentration, a mean function, a variance function and a parametric distributional family are required. The mean function is estimated as described in Section 7.0.

As described in Section 1.2, the gamma distribution is used for analytical methods that give only non-negative measurements, and the normal distribution is used for methods that can give both positive and negative measurements at low concentration. The gamma distribution is appropriate in the first case since it is the maximum entropy distribution (Kagan et al., 1973) in the class of continuous distributions which support only non-negative values and have a specified mean and variance (or specified mean and geometric mean). The normal distribution is the continuous maximum entropy distribution with specified mean and variance and support on the real line.

9.2 Prediction Variance

As described in Sections 6.0 and 8.0, models are constructed for both replicate variance and conditional MSE (of residuals from the means model) as functions of spiking concentration. For estimation of the LCMRL and HV-DL, what is really needed is a model for prediction variance as a function of spiking concentration. Three additional considerations allow construction of a robust yet conservative model for prediction variance.

The first is that, as discussed in Section 6.1, the variance at extremely low concentrations may be distorted by processes such as thresholding, smoothing, rounding and truncation. Therefore, a conservative prediction variance function should never give a zero prediction variance, not even at zero spiking concentration.

Because the constant in either the replicate variance (equation 8) or cMSE (equation 11) models may be zero, a parameter called minVar has been created. The minVar parameter is set to the constant *a* if it is nonzero. Otherwise it is set to the mean of the robust variances at the two lowest nonzero spiking levels. The minVar parameter is denoted by $\sigma^2_{\min}$ and $\tau^2_{\min}$ for the replicate variance and cMSE models, respectively.

The second consideration is that as noted in Section 8.1, although the replicate variance at each spiking level should always be smaller than the cMSE, it is possible for the replicate variance *model* estimate to be larger at some spiking concentrations than the cMSE *model* estimate. A conservative approach is to use the point-wise maximum of these functions.

The third consideration is that the prediction variance is associated in a sense to the regression (conditional means model). The prediction variance should be augmented with regression model uncertainty just as it is in ordinary least squares (OLS) regression (Draper and Smith, 1981). This makes intervals calculated using the prediction variance comparable to tolerance intervals.

These considerations guide the construction of the prediction variance function:

$$
\begin{aligned}
\bar{x} &= \tfrac{1}{n}\sum_{j=1}^{m} n_j x_j \\
f(x) &= 1 + \frac{1}{n} + \frac{(x-\bar{x})^2}{\sum_{j=1}^{m} n_j (x_j - \bar{x})^2} \\
\sigma^2(x) &= a_v + b_v x^{c_v} \\
\tau^2(x) &= a_e + b_e x^{c_e} \\
\sigma_{mm}^2 &= \max\left(\sigma_{\min}^2, \tau_{\min}^2\right) \\
\sigma_{pred}^2(x) &= \begin{cases} \max\left(\sigma^2(x), \tau^2(x)\right) \cdot f(x), & \text{if } \max(a_v, a_e) > 0 \\ \max\left(\sigma_{mm}^2, \sigma^2(x), \tau^2(x)\right) \cdot f(x), & \text{otherwise} \end{cases}
\end{aligned}
\qquad (12)
$$

Prediction variance estimates from this function are assumed to have $d = n - m - n_p$ degrees of freedom, where n_p is the maximum of the number of parameters fit in the variance model and the cMSE model.

9.3 Conditional Distributions

Maximum entropy is one method of characterizing probability distributions (Kagan et al., 1973). Often there is a unique distribution that maximizes entropy while satisfying conditions of allowable values for a random variable and constraints on specified moments (or on other functionals). This unique distribution satisfies the given conditions while implying absolutely no additional information. Maximum entropy distributions are very important when working with scientific problems that have physical constraints. They can honor those constraints without adding any additional unwarranted information.

The gamma distribution is the maximum entropy distribution in the class of continuous distributions with support (that is, allowed values) on the nonnegative real line and specified mean and geometric mean (Kagan et al., 1973). This meshes ideally with the model for low-level analytical error variance described by Rocke and Lorenzato (1995) and discussed in Section 6.1. It has both additive and multiplicative error components, corresponding to the linear and log-scale moment constraints in the gamma distribution.

It can easily be shown (Kagan et al., 1973) that in this case, the constraint on mean and geometric mean is equivalent to a constraint on mean and variance. Thus, for specified conditional mean and variance functions with positive range the principle of Maximum Entropy generates a family of gamma distributions. Therefore, the gamma distribution is used as the distributional model for response for methods which do not give negative results.

The normal distribution is the continuous maximum entropy distribution with specified mean and variance and support on the real line (that is, negative values are allowed). The normal distribution is therefore used as the distributional model for response for methods which can give negative results. In the case of the normal model for repeated measurements, the conditional

mean and prediction variance functions directly give the parameters for the repeated measurement model. Since the variance function estimates have an associated degrees of freedom *d* that will typically be less than 30, probability calculations are made using the *t*-distribution with *d* degrees of freedom.

In the case of the gamma distribution (used for methods which cannot give negative results), the conditional gamma distribution parameters are computed as

$$\begin{aligned}\alpha(x) &= \mu^2(x)\big/\sigma^2_{pred}(x)\\ \beta(x) &= \sigma^2_{pred}(x)\big/\mu(x).\end{aligned} \tag{13}$$

9.4 Search for LCMRL

An iterative search is conducted for the spiking level which satisfies the definition of the LCMRL. In the case of the normal distribution model, the LCMRL is the *x* which satisfies

$$\begin{aligned}&\Pr\left(0.5x \le Y < 1.5x \middle| \mu(x), \sigma^2_{pred}(x)\right) = T_d\left(\frac{1.5x-\mu(x)}{\sigma_{pred}(x)}\right) - T_d\left(\frac{0.5x-\mu(x)}{\sigma_{pred}(x)}\right) = \gamma_Q\\ &d = n - m - n_p\end{aligned} \tag{14}$$

where $T_d(\cdot)$ is the cumulative distribution function for the *t*-distribution with *d* degrees of freedom and $\gamma_Q = 0.99$.

In the case of the gamma distribution model, the LCMRL is the *x* which satisfies

$$\begin{aligned}&\Pr\left(0.5x \le Y < 1.5x \middle| \alpha(x), \beta(x)\right) =\\ &\quad F\left(1.5x \middle| \alpha(x), \beta(x)\right) - F\left(0.5x \middle| \alpha(x), \beta(x)\right) = \gamma_Q\end{aligned} \tag{15}$$

where $F\left(\cdot \middle| \alpha, \beta\right)$ is the gamma cumulative distribution function with parameters α and β.

Figure 9-1 below shows an example of the coverage probability for the MQO recovery interval (50%-150%) as a function of spiking level. Figure 9-2 below is an example of an LCMRL plot showing the raw data, the estimated average response curve, the recovery limits, a 99% prediction envelope for the response and the LCMRL. The text on the figures also gives the LCMRL. These figures were created using the LCMRL calculator software application produced for laboratory use.

In this example, the spiking range is relatively narrow and the replicate variance and conditional mean squared error are relatively constant across the range. Figure 9-3 below shows an example of an LCMRL plot for an analyte with error (replicate variance and conditional MSE) increasing with spiking concentration.

Figure 9-1: Example Coverage Probability for MQO Recovery Limits

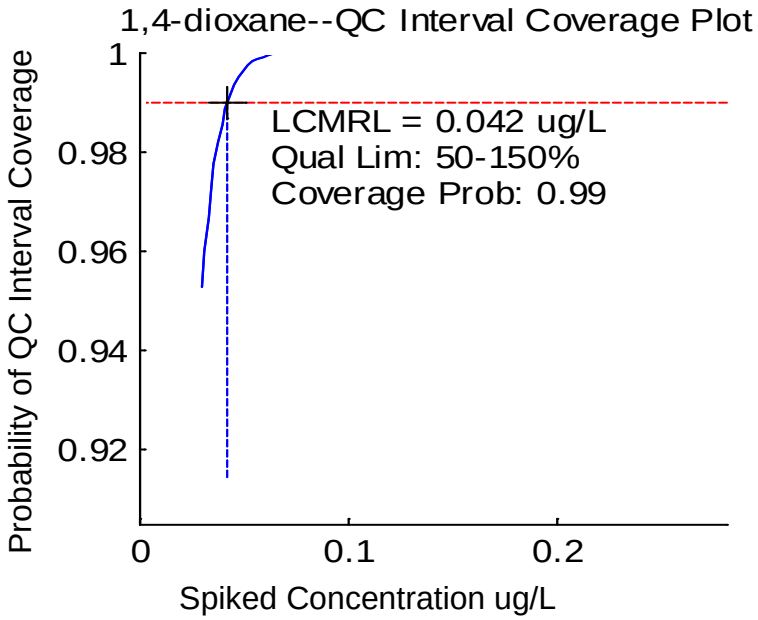


Figure 9-2: Example LCMRL Plot with Constant Error Variance

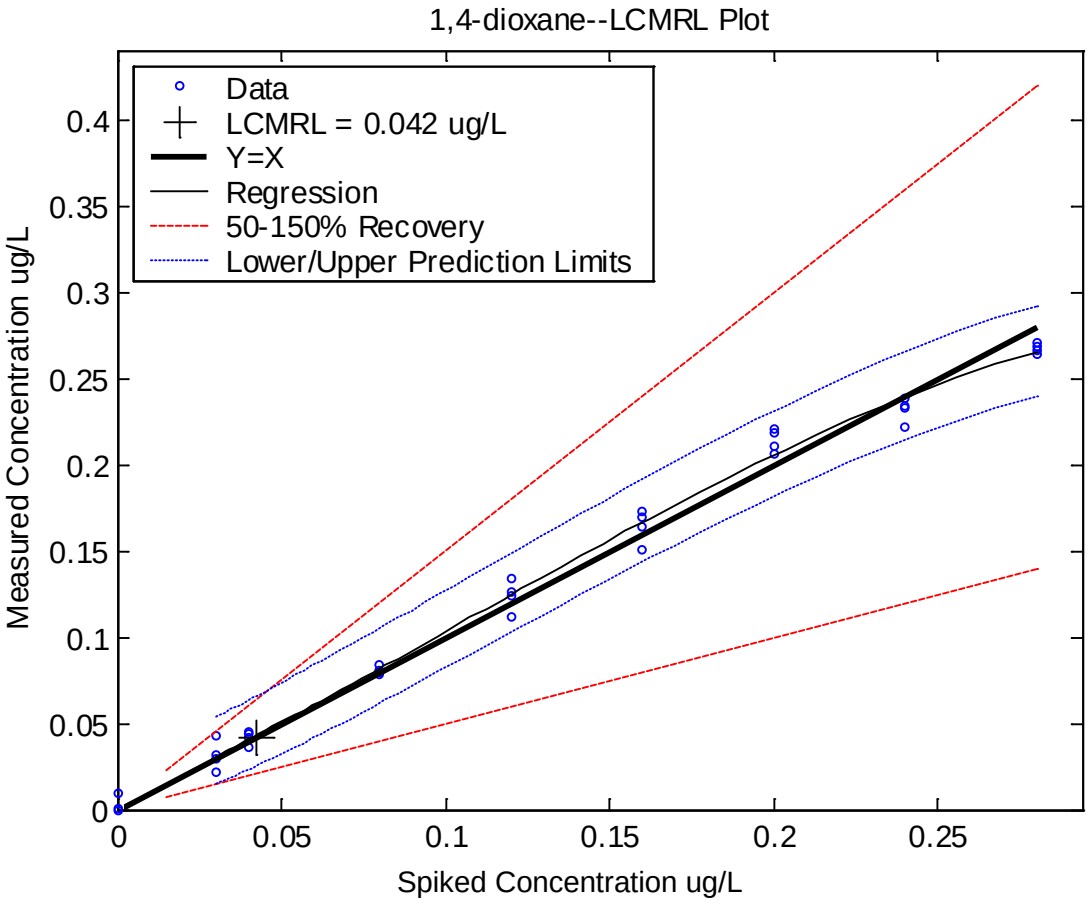


Figure 9-3: Example LCMRL Plot with Increasing Error Variance

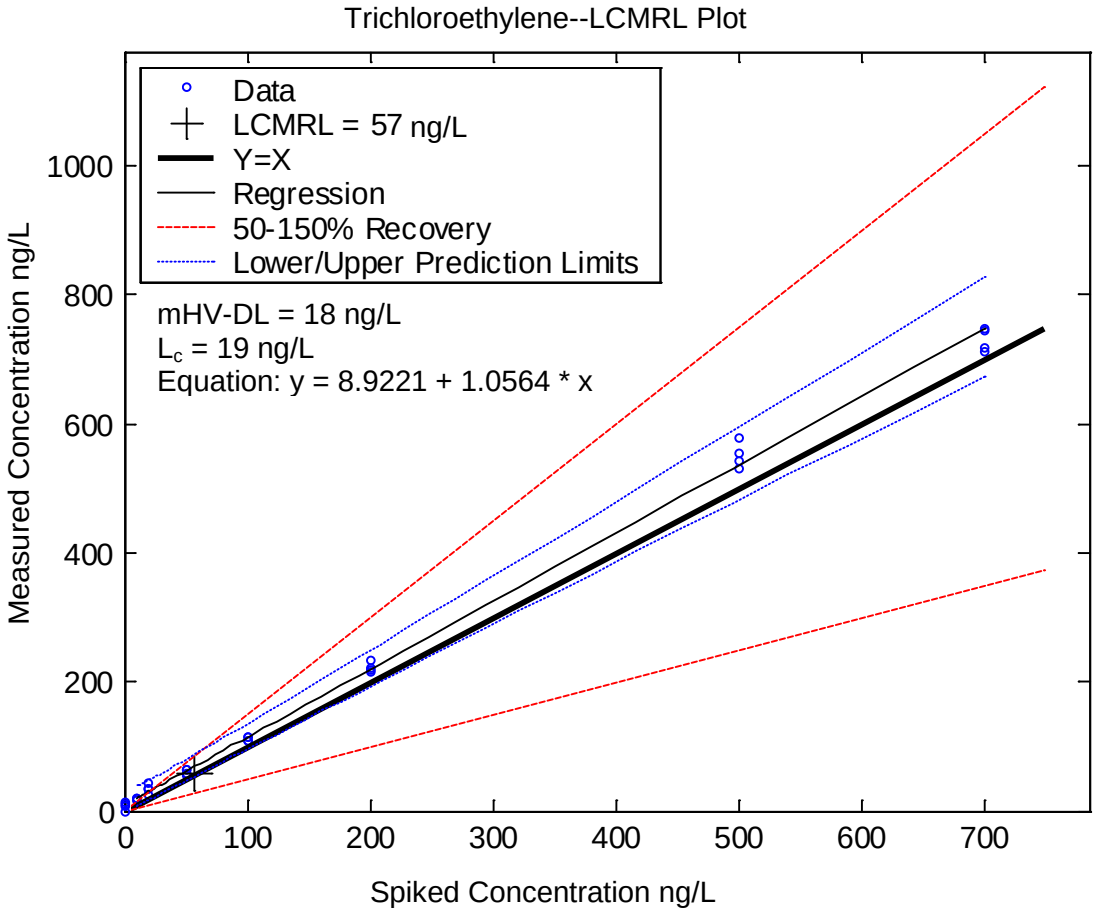
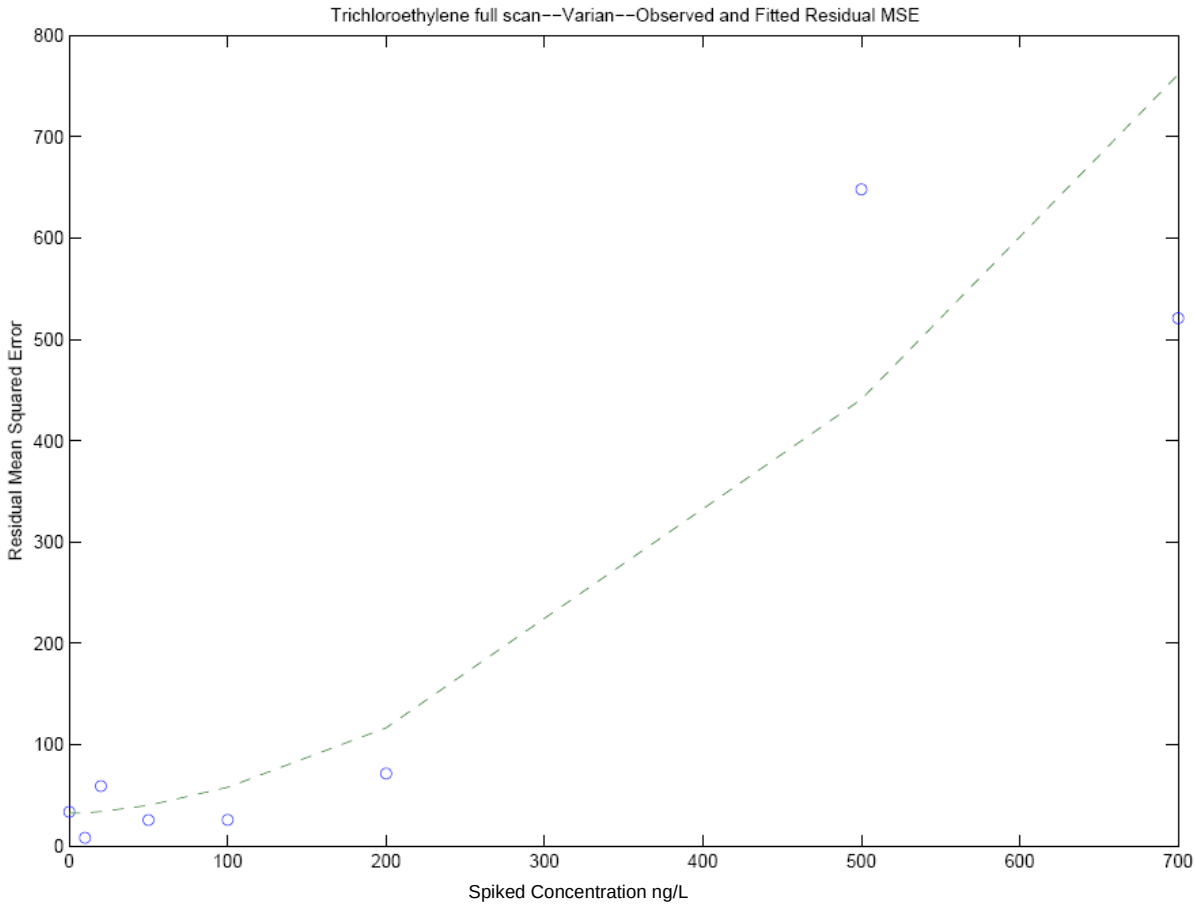


Figure 9-4 below shows the increasing MSE as a function of spiking level. The points are the robust estimates of MSE. The line is an approximation to the constant + power model fitted to the cMSE.

Figure 9-4: Conditional MSE and fitted model showing increasing error with concentration



10.0 Estimating the Modified Hubaux-Vos DL

The proposed methodology differs from and improves on the original Hubaux-Vos methodology (Hubaux and Vos, 1970). First, the proposed method accommodates nonconstant variance in a very robust manner. Hubaux and Vos were aware of this consideration but did not include it in their methodology. Secondly, in cases where the instrument response cannot be negative, the proposed methodology uses the gamma distribution (which is the Maximum Entropy distribution in this situation) to model the response distribution.

10.1 Estimating the Critical Level

L_c was defined by Currie (1968) as an upper percentile of the distribution of repeated measurements at zero concentration. For Currie, L_c represents an *a posteriori* decision limit for detection. As noted in Section 6.1, although in many situations it may be impossible to sample this distribution directly in an accurate way, it is possible to estimate using regression methods. This was the general approach of Hubaux and Vos (1970).

As is the case for the LCMRL (Section 9.1), the methodology used to estimate L_c in the mHV-DL method distinguishes two cases. In the first case, in which blank and low-level measurements *can* give negative instrument responses, the normal distribution is used to model the response distribution. Then L_c is estimated as

$$\begin{aligned} L_C &= \mu(0) + \sigma_{pred}(0) \cdot t_d(\gamma_C) \\ d &= n - m - n_p \end{aligned} \tag{16}$$

where $t_d(\gamma_C)$ is the γ_C-quantile of the *t*-distribution with *d* degrees of freedom (see Section 9.2), and $\gamma_C = 0.95$.

In the second case, in which blank and low-level measurements *cannot* give negative instrument responses, the gamma distribution is used to model the response distribution.
Then, assuming that $\mu(0) > 0$, L_c is estimated as

$$L_C = F^{-1}\left(\gamma_C \middle| \alpha(0), \beta(0)\right) \tag{17}$$

where $F^{-1}\left(\cdot \middle| \alpha, \beta\right)$ is the quantile function of the gamma distribution with parameters α and β.

However, in the case that $\mu(0) = 0$, the gamma distribution is not tenable. In this case, for want of a better alternative, a half-*t* distribution is used. Then L_c is estimated as

$$L_C = t_d\left(1 - \tfrac{1-\gamma_C}{2}\right) \cdot \sigma_{pred}(0) \tag{18}$$

where $t_d(\cdot)$ is the quantile function of the *t* distribution with *d* degrees of freedom.

10.2 Search for mHV-DL

An iterative search is conducted for the spiking level which satisfies the definition of the mHV-DL. In the case of the normal distribution model, the mHV-DL is the *x* which satisfies

$$\Pr\left(Y \geq L_C \middle| \mu(x), \sigma^2_{pred}(x)\right) = T_d\left(\frac{L_C - \mu(x)}{\sigma_{pred}(x)}\right) = \gamma_D \tag{19}$$

where $T_d(\cdot)$ is the cumulative distribution function for the *t*-distribution with *d* degrees of freedom (see Sections 7.2 and 7.3), and $\gamma_D = 0.95$.

In the case of the gamma distribution model, the mHV-DL is the *x* which satisfies

$$\Pr\left(Y \geq L_C \middle| \alpha(x), \beta(x)\right) = 1 - F\left(L_C \middle| \alpha(x), \beta(x)\right) = \gamma_D \quad (20)$$

where $\alpha(x)$ and $\beta(x)$ are defined in equation 19, and $F\left(\cdot \middle| \alpha, \beta\right)$ is the gamma cumulative distribution function with parameters α and β.

For the data shown in Figure 9-2, the mHV-DL and L_c values are 0.026 ug/L and 0.016 ug/L, respectively. In this case, which is typical, the mHV-DL is higher than L_c, as one would intuitively expect. In the case of the TCE data used for Figure 9-3, the reverse is true. The mHV-DL and L_c values are 0.018 ug/L and 0.019 ug/L (both are converted from ng/L, as presented in Figure 9-3), respectively.

Although this seems odd at first glance, in fact, the mHV-DL is in the domain of spiking concentrations while L_c is in the measurement domain. Systematic low level bias in the measurements can cause this apparent anomaly. Even if a calibration does not show a statistical significant intercept term, because the LCMRL study involves much more data and because it includes preparation steps, it tends to show low level bias and nonlinearity not evident in the calibration data.

Let x_c denote the spiking concentration at which the expected response equals L_c. Then it is always true that x_c < mHV-DL. In fact, x_c is a very important quantity, since it is the natural censoring point to be used in data analysis for nondetects (measurements which are less than or equal to L_c). In the case of the TCE data used for Figure 9-3, the intercept for the mean response model, at about 0.009 ug/L, is about half L_c (0.019 ug/L), which is a significant fraction of L_c. Remember that L_c is the estimated 95th percentile of the response distribution at zero spiking concentration. To get a response greater than 0.019 ug/L 95% of the time, we estimate that we must spike at 0.018 ug/L (the mHV-DL). Although it may seem odd, this situation can be explained by the positive intercept of the mean response model being "not small" relative to L_c.

11.0 Summary, Conclusions and Recommendations

The proposed LCMRL procedure described in this report provides a statistically robust method of estimating the lowest concentration at which a laboratory can reliably achieve the MQO of 50%-150% recovery. Although the procedure is complex and computationally intensive, implementing it in user-friendly software in the public domain and freely available over the internet makes it easy for laboratories to use.

Furthermore, a robust method of estimating the LCMRL makes it possible, through Bayesian bootstrap resampling, to estimate by simulation the statistical distribution of LCMRL, mHV-DL and L_c values that would be generated by a randomly selected 'experienced drinking water laboratory' upon repeated execution of the LCMRL study design. This will enable the development of scientifically defensible MRL values for guidance and regulatory use.

Coupling the mHV-DL procedure with the modeling used for the LCMRL computation improves the estimation of L_c and the detection limit. Improvement of the estimation of L_c will improve statistical inference regarding low-level data sets. Improvement of the estimation of the

detection limit will improve the accuracy of assessment of risk modification due to regulatory actions.

These developments have great potential value for regulators, the regulated community, risk assessors and data analysts, but especially for laboratories. Use of these methods in the future should result in much better understanding and control of measurement data quality. In turn, this should lead ultimately to better data quality and decision-making.

The author has several recommendations for future work related to the LCMRL/Hubaux-Vos methodology. The first is the further characterization and potential improvement of the statistical methodologies proposed here, particularly viewing them as a combined procedure rather than isolated methodologies. Secondly is development of optimal design methodology (based on the statistical characteristics of the methodology) and development of more definitive recommendations for the design of LCMRL/DL studies.

The third recommendation is development or improvement of procedures for verification of MRL capability and ongoing monitoring of detection and quantitation capabilities in the context of a routine laboratory quality control program. The fourth recommendation is the development and release in the public domain of an R library (R Development Core Team, 2007) for the LCMRL/Hubaux-Vos methodology, which will facilitate the use and further development of this methodology in the research communities of statistics, chemometrics, environmental analytical chemistry and forensic and clinical chemistry.

12.0 References

Currie, L.A. (1968). "Limits for Qualitative Detection and Quantitative Determination." *Analytical Chemistry*, Vol. 40, pp. 586-593.

Dobson, A.J. (1983). *Introduction to Statistical Modeling*. Chapman and Hall, New York.

Draper, N. and H. Smith (1981). *Applied Regression Analysis: 2nd Edition*. John Wiley and Sons, New York.

Glaser, J.A., D.L. Foerst, G.D. McKee, S.A. Quane, and W.L. Budde (1981). "Trace Analyses for Wastewaters." *Environmental Science and Technology*. Vol. 15, pp. 1426-1435.

Hoaglin, D.C., F. Mosteller, and J.W. Tukey (1983). *Understanding Robust and Exploratory Data Analysis.* John Wiley, New York.

Horn, P.S. (1988). "A Biweight Prediction Interval for Random Samples." *Journal of the American Statistical Association*. Vol. 83, No. 401. (Mar., 1988), pp. 249-256.

Hubaux, A. and G. Vos (1970). "Decision and Detection Limits for Linear Calibration Curves." *Analytical Chemistry*. Vol. 42, No. 8, pp. 849-855.

Jaynes, E. (1983). *Papers on Probability, Statistics and Statistical Physics. A reprint collection*. R. D. Rosenkrantz (Ed.). Reidel. Dordrecht, Germany.

Kagan, A.M.; Y.V. Linnik, and C.R. Rao (1973). *Characterization Problems in Mathematical Statistics*. John Wiley. New York. 499 pp.

Kelley, T.L. (1921). "A New Measure of Dispersion." *Quarterly Publications of the American Statistical Association*. Vol. 17, No. 134. pp. 743-749.

Lax, D.A. (1985). "Robust estimators of scale: Finite-sample performance in long-tailed symmetric distributions," *Journal of the American Statistical Association.* Vol. 80, pp. 736-741.

McCullagh, P. and J.A. Nelder (1989). *Generalized Linear Models: 2nd Edition*. Chapman & Hall, London.

Nelder, J.A. and R. Mead (1965). "A Simplex Method for Function Minimization", *Computer Journal*. Vol. 7, pp. 308-313.

R Development Core Team (2007). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Randles, R.H. and D.A. Wolfe (1991). *Introduction to the Theory of Nonparametric Statistics*. Krieger Publishing, Malabar, Florida.

Rocke, D.M. and S. Lorenzato (1995). "A Two-Component Model for Measurement Error in Analytical Chemistry." *Technometrics*. Vol. 37, No. 2, pp. 176-184.

Rubin, D.B. (1981). "The Bayesian Bootstrap." *Annals of Statistics*, 9, pp. 130-134.

Thiel, H. and D.G. Fiebig (1984). *Exploiting Continuity: Maximum Entropy Estimation of Continuous Distributions*. Ballinger Publishing Co. Cambridge, MA. 246 pp.

USEPA (1987). "National Primary Drinking Water Regulations-Synthetic Organic Chemicals; Monitoring for Unregulated Contaminants; Final Rule." *Federal Register*. Vol. 52, No. 130. p. 25690, July 8, 1987.

USEPA (1999). "Revisions to the Unregulated Contaminant Monitoring Regulation for Public Water Systems; Final Rule." *Federal Register*. Vol. 64, No. 180. p. 50556, September 17, 1999.

USEPA (2004). *Statistical Protocol for the Determination of the Single-Laboratory Lowest Concentration Minimum Reporting Level (LCMRL) and Validation of Laboratory Performance at or Below the Minimum Reporting Level (MRL).* USEPA, Office of Ground Water and Drinking Water, Standards and Risk Management Division, Technical Support Center. EPA 815-R-05-006. November 2004.

Appendix A: Computation of MRL

The computations carried out for establishing the Minimum Reporting Level (MRL) are programmed in the R language. The Matlab code for the LCMRL calculator was first translated into R and then amended to also produce estimates of the MRL estimate for all analytes appropriate for use in the LCMRL calculator. There are two reasons for programming the MRL in R; the first being that the recoding of the Matlab code provided a validation check of the LCMRL code, and secondly it provided an interactive, non-compiled program that could be used internally by EPA to calculate MRL estimates and associated graphics.

The MRL is determined using a Bayesian bootstrap (BB) (Rubin, 1981) of the LCMRL estimator using the LCMRL study data from each of several experienced drinking water laboratories. The BB replicates that were generated from each laboratory's data serve to estimate the distribution of estimated LCMRL values that each laboratory might generate on repeated performance of the LCMRL study. The distribution of pooled BB replicates, generated from LCMRL study data, can be used to approximate the distribution of estimated LCMRL values which might be generated from the population of experienced drinking water laboratories.

The MRL is calculated in three steps whenever there are three or more laboratories providing data with valid LCMRLs or calculated LCMRLs that are below the lowest non-zero spiking level. In the first step, 200 BB LCMRL replicates are calculated for each laboratory data set. In the second step a predicted distribution of some unknown and yet to be observed laboratory is built from the population of replicate laboratory LCMRLs using a random effects model. In the third and last step the MRL is taken to be the upper 95% one-sided confidence interval on the 75^{th} percentile of the predicted distribution referred to as the 95-75 upper tolerance limit (95-75 UTL).

A description of the procedure is given in the following three steps.

Step 1: Bayesian Bootstrap of the Laboratory data

When three or more laboratories have either a valid LCMRL or an estimated LCMRL that is reported to be less than the lowest reported non-zero spiking level a BB is carried out on each laboratory's data to produce 200 replicate LCMRL estimates, l_{kr} . Here k is the number of laboratories ($k \geq 3$), and r is the number of replicates (200),. The BB is implemented by selecting random Dirichlet weights from the uniform distribution on the $n_i - 1$ dimensional simplex, where n_i is the number of measurements at the i^{th} spiking level. These weights are associated with the original replicate measurements within each of the i spiking levels of the laboratory data. The LCMRL procedure is then rerun using the Dirichlet weights to compute a weighted LCMRL. This is done repeatedly to generate the sample of BB replicate LCMRL estimates.
Using Efron's bootstrap on a small data set, like the set of replicates at each spiking level, is not appropriate. Statistics computed using Efron's bootstrap are essentially calculated as randomly weighted statistics where the set of possible weights for each observation is $\left\{0, \frac{1}{n}, \frac{2}{n}, \frac{3}{n}, \ldots, \frac{n-1}{n}, 1\right\}$ with the additional constraint that all of the weights sum to 1. When n is small, the number of possible weights and therefore the number of values of any statistic that could be computed from the data given those weights, is small. Therefore our approach using the Bayesian Bootstrap with Dirichlet weights allows an extremely large number of weighted LCMRL values to be computed even from a relatively small data set.

The portions of the LCMRL calculation that use robust weights are modified to use a product weight of the form

$$w_{ij}^{*} = w_{ij}^{D} * w_{ij}^{RW},\ j = 1,\ldots,n_i$$

$$w_{ij} = w_{ij}^{*} / \sum_{j=1}^{n_i} w_{ij}^{*},\ j = 1,\ldots,n_i$$

where w_{ij}^{D} is the Dirichlet weight and w_{ij}^{RW} is either the robust Huber or Tukey bisquare weight. The weights are then normalized to produce the final weights w_{ij}.

Step 2: Estimating the Predictive Distribution

The predictive distribution of an unknown and unobserved laboratory from the random sample of laboratories reporting data is estimated in three steps as follows

1. **Power Transform of BB replicate Laboratory Data**

   The BB replicate results are transformed to approximate normality using the power transform (Box and Cox, 1964). The transformed replicate LCMRL values, l_{kr}^{*}, are calculated as follows where λ is the transformation parameter

   $$l_{kr}^{*} = \begin{cases} l_{kr}^{\lambda},\ k = 1,\ldots K,\ r = 1,\ldots,200 \text{ if } \lambda > 0 \\ \ln(l_{kr}^{\lambda}), k = 1,\ldots K,\ r = 1,\ldots,200 \text{ if } \lambda = 0 \end{cases}$$

   where K is the number of laboratories. The transformation parameter λ is constrained to be non-negative.

2. **The values of the predictive distribution are calculated according to the following steps:**

   a. **Calculate Initial Resistant Location and Scale Estimates**

      Location and scale parameters are estimated for the pooled set of transformed replicate values. These values are used as initial estimates in generating one-sided Tukey bisquare weights to down-weight large estimates in the pooled transformed replicate LCMRL values l_{kr}^{*}. The median of the l_{kr}^{*} is used for the initial resistant location estimate and the median absolute deviation from the median (MAD) of the l_{kr}^{*} is used as the initial resistant scale estimate.

   b. **Calculate One-Sided Tukey Bisquare Weights for the pooled LCMRL replicate values**

      The one sided Tukey bisquare weights are calculated using a tuning constant, c_{BW}, of 6. The initial starting values used are the median, $T_{BW}^{(0)}$, and MAD from "Step 2a" above. If any of the weights are zero, the weights are recalculated with a tuning constant of 9. The weights are iteratively calculated at the t^{th} stage according to the following method

$$u_{kr}^{(t)} = \begin{cases} \frac{l_{kr}^{*} - T_{BW}^{(t-1)}}{c_{BW} AD}, k = 1,\ldots,K, r = 1,\ldots,200, \text{ if } l_{kr}^{*} > T_{BW}^{(t-1)} \\ 0, \text{ if } l_{kr}^{*} \le T_{BW}^{(t-1)} \end{cases}$$

$$w_{kr}^{BW(t)} = \begin{cases} \left(1 - \left(u_{kr}^{(t)}\right)^2\right)^2, & \text{if } u_{kr}^{(t)} < 1 \\ 0, & \text{if } u_{kr}^{(t)} \ge 1 \end{cases}, k = 1,\ldots,K, r = 1,\ldots,200$$

$$T_{BW}^{(t)} = \left(\sum\nolimits_{kr} w_{kr}^{BW(t)} l_{kr}\right) \Big/ \sum\nolimits_{kr} w_{kr}^{BW(t)}$$

until a relative convergence of 1E-6 is reached for $T_{BW}^{(t)}$. After convergence the final one-sided robust weights are calculated as the normalized weights

$$w_{kr}^{BW} = \frac{w_{kr}^{BW(t)}}{\sum_{kr} w_{kr}^{BW(t)}}$$

c. Calculate Robust means and Variances for each of the *K* laboratory replicate LCMRL sets

Robust location and scale estimates are calculated by laboratory for the transformed replicate LCMRL values, l_{kr}^{*}. These values are used in estimating the predicted laboratories variance on the transformed scale. For the k^{th} laboratory, an estimate of the mean, m_k^{*}, and variance , s_k^{*2}, are calculated using the blended Huber and Tukey biweight methodology used in the LCRML routine except that a tuning constant of 6 is used instead of 9.

d. Estimate the variance of the predicted distribution

Robust location and scale estimates are estimated by laboratory for the transformed replicate LCMRL values. The predicted variance, s_{pred}^2, is then estimated as

$$m^{*} = \sum\nolimits_{k=1}^{K} \left(\sum\nolimits_{r=1}^{200} w_{kr}\right) m_k^{*}$$

$$s_w^2 = \sum\nolimits_{k=1}^{K} \left(\sum\nolimits_{r=1}^{200} w_{kr}\right) s_k^{*2}$$

$$s_b^2 = \frac{\sum_{k=1}^{K} \left(\sum_{r=1}^{200} w_{kr}\right) \left(m_k^{*} - m^{*}\right)^2}{1 - \sum_{k=1}^{K} \left(\sum_{r=1}^{200} w_{kr}\right)^2}$$

$$s_{pred}^2 = \left(1 + 1/K\right) s_b^2 + s_w^2$$

e. Estimate the values of the predicted distribution

The values of the predicted distribution are generated as follows. First the transformed values are generated as

$$l^*_{pred,kr} = m^* + s_{pred}(l^*_{kr} - m^*_k)/s^*_k,\ k = 1,\ldots,K,\ r = 1,\ldots,200$$

These values are then transformed back to the original scale as

$$l_{pred,kr} = \begin{cases} \exp(\ln(l^*_{pred,kr})/\lambda), \text{ if } \lambda > 0 \\ \exp(l^*_{pred,kr}), \text{ if } \lambda = 0 \end{cases}$$

Step 3: Estimate the MRL as the 95-75 UTL of the Predictive Distribution

The MRL is then estimated as the 95-75 UTL from the values of the predicted distribution using a weighted version of the Guttman non-parametric procedure (Guttman, 1970) using the weights w_{kr}.

References

Box, G. E. P. and D.R. Cox (1964). "An Analysis of Transformations," *Journal of the Royal Statistical Society*, Series B. Vol 26, pp. 211-46.

Guttman, I. (1970). *Statistical Tolerance Regions: Classical and Bayesian***.** Hafner Press, Darien, CT.

Rubin, D. B. (1981). "The Bayesian Bootstrap." *Annals of Statistics*, 9, pp. 130-134.