

Not to be quoted without prior reference to the authors

Fisheries Research Services Collaborative Report No 04/04

**THE APPLICATION OF THE COFINO MODEL TO EVALUATE
LABORATORY PERFORMANCE STUDY DATA USING
THE BANDWIDTH ESTIMATOR**

D E Wells, W P Cofino and J A Scurfield

June 2004

Fisheries Research Services
Marine Laboratory
Victoria Road
Aberdeen AB11 9DB

THE APPLICATION OF THE COFINO MODEL TO EVALUATE LABORATORY PERFORMANCE STUDY DATA USING THE BANDWIDTH ESTIMATOR

David E Wells¹, Wim P Cofino² and Judith A Scurfield¹

¹QUASIMEME Project, Fisheries Research Services,
Marine Laboratory, PO Box 101, 375, Victoria Road,
Aberdeen, AB11 9DB UK

²Wageningen University, Environmental Sciences, Sub-Department of Water Research,
Hydrology and Quantitative Water Management Group, Nieuwe Kanaal 11,
6708 PA Wageningen, The Netherlands

PREFACE

This handbook describes the application of the Cofino model for the evaluation of Laboratory Performance (LP) data. It covers the development of the model and the reasons for this alternative approach to data analysis.

An overview of the mathematical basis for the model is described with the details being found in publications elsewhere. The model is validated with data of known characteristics including normal and bimodal distributions, and published data.

A wide range of examples have been taken from the QUASIMEME LP studies, the Food Standards Agency (FSA) Genetically Modified Organisms (GMO) Proficiency Testing Scheme and the International Atomic Energy Agency (IAEA) interlaboratory studies.

This handbook currently only describes the model and its application to LP data. The model, however, is applicable to any data set for which the population characteristics are required. The future development of this handbook will include worked examples of the actual program, the documentation of the programming code and the compiled programs¹.

SUMMARY

The Cofino model has been developed and tested for use in the determination of population characteristics specifically, but not exclusively for laboratory performance studies.

The model identifies clusters of values within a dataset that exhibit a high level of agreement and it calculates the mean, standard deviation (s.d.) and the percentage of data from the whole dataset associated with each cluster. The summary statistics are available for each cluster of data.

The model can be used directly with tailing and skewed data, datasets containing extreme values (outliers) and bimodal (multi-modal) distributions. In the case of bimodal data, the model will identify each mode and provide an estimate of the mean, standard deviation and percentage of data associated with each mode.

¹The model is programmed in MATLAB v6 with the statistical toolbox. Copies of the programs are available from the authors. However, the compiled version will be built as a standalone executable with the associated libraries that will not require the user to have MATLAB installed before use.

The cluster that represents the greatest percentage is equal to the first mode of the dataset. Other clusters may also represent modes. The population characteristics of the first mode are a good estimate for those of the whole data.

No selection or preparation of data is necessary, either by using subjective boundaries or outlier testing prior to using the Cofino model.

The model includes left censored values (LCVs) (*less than values*) in the datasets and has been tested on datasets with up to 50% LCVs.

The Cofino model uses probability density functions for each observation as a starting point. Generally, a normal distribution is used which implies that for each observation the value and the s.d._{within} have to be specified.

The Cofino model has four methods of operation to establish a value for the s.d._{within} that is used as an input to the model.

- i) The Normal Distribution Assumption (NDA) model which uses the whole dataset to reconstruct the s.d._{within} for use within the model.
- ii) The BandWidth Estimator (BWE) uses a selected *heart-cut* of data from the dataset to reconstruct the s.d._{within}. The BWE is based on the group of values that are in good agreement.
- iii) The individual laboratory replicate measurements ($r > 5$) to calculate the s.d._{within}.
- iv) The average standard deviation of laboratory replicate measurements, where the number of replicates < 5 or where there is poor agreement for the individual s.d._{within}.

The model works well **without** the need to have a measure of the laboratory uncertainties (Individual s.d._{within}) by using the BWE developed for this purpose. The model, using the BWE, provides well-defined structural details and can be used in conjunction with information on methodology to identify preferred analytical techniques.

The detailed, graphical information from the model provide:

- i) A ranked overview of the means and the s.d. of each data set and the expectation values. This model includes normal and left censored values.
- ii) A plot of the population measurement functions (PMFs).
- iii) A plot of the population density functions.
- iv) 2D and 3D plots of the first 2 and 3 expectation values of PMF_1 , PMF_2 , & PMF_3 . These plots are useful for detecting differences in methodology.
- v) Matrix overlap (Kilt) plot. This colour density plot is very sensitive to identifying structure of data, especially modality, where it is not so clear in other plots. Bimodality can be detected with the Kilt plot even when both modes have equal density.
- vi) Z score plot for reviewing performance against targets.

The summary statistics provide:

- i) An estimate of the mean for each mode or cluster.
- ii) The standard deviation of each mode or cluster.
- iii) The percentage of data associated with each mode or cluster.
- iv) These data are used concurrently for the NDA and BWE model or the NDA and the laboratory s.d._{within} model.
- v) The Kolmogorov-Smirnov test for normality.
- vi) The Student's t test- a preliminary check for bimodality between the first two modes.

The model has been fully tested with:

- i) Randomly generated normal distributions with the number of observations (NObs) 10-200, relative s.d. 2-30% to compare performance with conventional statistics.
- ii) Normal distributions overlaid to construct bimodal data. The sensitivity to bimodality with respect to the distance and the relative amplitude of the modes.
- iii) A comparison with the Kernel Density Estimator, robust statistics and bootstrapping.

The evaluation of 2055 datasets taken from the QUASIMEME Laboratory Performance Studies.

- iv) The inclusion of left censored values.

The Cofino model has been developed using the MATLAB programming language. Copies of the code and associated information are available from the authors. A compiled version is available for use without the need to obtain the MATLAB programs. The compiled versions cannot be amended.

Currently, no other model provides similar comprehensive summary statistics for each mode in a dataset and extensive graphical output describing the structural information in the data. These data and information can be obtained in a single computational operation without any manipulation, trimming or subjective elimination, including left censored and missing values.

1. INTRODUCTION

Laboratory Performance (LP) study or Proficiency Test (PT) data has historically been assessed using ISO 5725 [ISO 1994] to obtain an estimate of the mean and the uncertainty of the measurement. They have been more recently assessed using robust statistics [AMC 1998a & b, Cofino & Wells 1994, Wells & Cofino 1997, de Boer *et al* 2000].

ISO 5725 [1994], which originated from work carried out by Youden and Steiner [1975], provides a series of outlier tests for LP study data. This standard was primarily developed for method performance studies where each laboratory was required to use the same protocol. This approach assumes an equal *within-laboratory* variance and a normal distribution of values. Application of ISO 5725 to LP studies is questionable, as the requirement for normality is frequently violated and *within-laboratory* variance is often variable since different analytical methods are used. Transformation of data can sometimes overcome this deviation from normality and non-equal *within-laboratory* variances. However the use of ISO 5725 is not straightforward when outlier tests are applied to such data. Data from an LP studies can be skewed, either positively or negatively or bi- or multi-modal, depending on the nature of the measurement.

In principle, laboratories should participate in these LP studies only when they have fully validated and statistically controlled methodologies and under these circumstances the data from these exercises are normally distributed. Since many LP studies are based on world-wide participation the standards and methodologies used can be very different in the different countries. This often gives rise to systematic differences in the data reported. Participation in these schemes is also fluid, with new participants taking 2-5 studies to establish themselves and improve to a satisfactory level of performance. Experienced laboratories can also vary in performance due to new staff requiring training, new methodologies and instrumentation. Because of the heterogeneous nature of these data a number of statistical tools have been developed recently to evaluate the information and provide each participant with an objective and accurate assessment of their data.

The evaluation of the LP studies data should provide:

- i) The most robust and reliable estimate of the concentration of the determinand.
- ii) The best estimate of the variance of the data, without the influence of the extreme values, many of which are random in origin.
- iii) An assessment of the individual laboratory performance using the above estimators.

For a meaningful assessment the assigned value should preferably be very close to the true value. In practise, and in the absence of other information, the consensus value is taken as the best estimate of the true value. This assumption seems to be reasonable when the data are normally distributed. With complex distributions, the consensus value is obtained with data from the '*best performing laboratories*'. The use of outlier tests is an example of this approach, however the key question is: "which approach should be used to identify the *best performing laboratories* from the entire dataset?"

Robust statistics [Huber 1972, Hoaglen *et al.*, 1983, AMC 1989a & b, AMC 2001] have been applied to data from interlaboratory studies in an attempt to overcome the difficulties associated with non-normally distributed data. With this technique, extreme values are not discarded as outliers, but *downweighted* to minimise the effect on the dataset as a whole. This approach attempts to describe the data that are in good agreement i.e. from the *best performing laboratories*. The application of robust statistics works well with datasets that

have up to effectively 5-7% of extreme values, even when the data are highly skewed [Wells & Cofino 1997]. However, the robust means are affected with a larger proportion of extreme values and/or when the evaluation is on a small number (*ca* < 10) of values. The main cause in the breakdown of the robust values occurs when the resultant group of *downweighted* values aggregate to form their own cluster. This influences the magnitude of both the robust mean and robust standard deviation (s.d.) when there are a relatively large number of tailing values (Fig. 1).

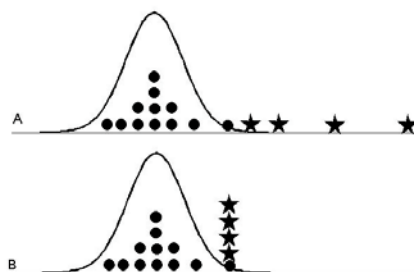


Figure 1. **A)** The distribution of hypothetical data with a group with relatively good agreement forming a normal distribution, shown by the curve, and a group of outliers.

If all of the outliers could be easily identified and removed then the mean and standard deviation would reasonably describe the group of *best performing* laboratories.

B) Robust Statistics attempts to overcome the effect that the outliers have on the group as a whole. However, the effect of such a relatively large number of outliers is not completely overcome by using Robust Statistics. Robust Statistics down-weights the large number of extreme values, but creates a cluster of data at the down-weight value affecting the mean and standard deviation of the group of *best performing* laboratories.

From 1992 until 2000 QUASIMEME² (Quality Assurance of Information in Marine Environmental Monitoring in Europe) used the robust statistical approach to assess the data from the LP studies [Wells & Cofino 1997]. The extreme values reported by participating laboratories can, in many cases, be traced back to calculation or transcription errors, use of incorrect units, wrong calibration, analytical errors or contamination. In practice many of these errors can be identified by inspection of the data and with good contact with the participants. However, elimination of extreme values brings a subjective element into the evaluation, which the robust statistics intended to avoid.

A new approach has been developed using the Cofino model [Cofino *et al.*, 2000] that solves many of these difficulties when applied to LP studies where the data is non-normally distributed. However, the scope of the model is much broader and can be applied to data where information on the population characteristics is required. The Cofino model borrows the concept of wavefunctions from quantum mechanics and applies these utilising the power of matrix algebra. The estimators of the mean and the variance that are derived from this

²QUASIMEME was funded by the European Union (1992-1995) for marine laboratories in Europe submitting data to international monitoring programmes such as the Oslo & Paris Commission (OSPAR), the Helsinki Commission (HELCOM) and the Barcelona Convention (MEDPOL). Since 1995 QUASIMEME has extended a worldwide programme and is funded by the subscribing laboratories

model are considerably less sensitive to asymmetric, tailing, data. Both the numerical and graphical tools developed with the output of this model can be used to provide information on the distribution, modality and homogeneity of the data.

The Cofino model attributes a probability density function to each datapoint. Usually, normal distributions are used. Consequently, in addition to the measured values, an estimate of the *within-laboratory* standard deviation is required. In the original paper, two approaches are described to implement the model. In the first approach, the *within-laboratory* standard deviations are supplied, for instance, by requesting laboratories to submit an estimate. The second approach, denoted as the normal distribution approximation (NDA), constructs a *within-laboratory* standard deviation from the measured values in such a way that the mean and standard deviation of normally distributed data are reproduced.

The main constraint in utilising the full potential of the first method is the availability of reliable estimates of the *within-laboratory* uncertainty of the measurements.

However, the Cofino model can be used very effectively in the absence of *within-laboratory* uncertainties. The NDA approach is based on the premise that the population characteristics of a normal distribution can be obtained when each data point is given an uncertainty calculated from the whole dataset. In most cases the NDA approach works well, but an improvement on this NDA model is desirable for datasets where the deviation from normality is large, due to bimodality or tailing values.

In this paper we have developed the use of a BandWidth Estimator (BWE). The BWE is a refinement of the NDA approach. The BWE and NDA are equivalent for normally distributed datasets. The BWE approach, however, provides better results for datasets with complicated underlying distributions owing to an improved procedure to construct the *within-laboratory* standard deviation. The development of the BWE is explained.

In this paper, the Cofino BWE model is applied to difficult datasets and compared with the Kernel Density Estimator (KDE) approach [Wand & Jones 1995, Lowthian and Thompson 2002, AMC 2002] as well as robust and the ISO 5725 statistics. Finally, these different models are compared and contrasted with the QUASIMEME LP studies for the 6 years (1996-2001) for all of the mandatory parameters for the OSPAR³ and HELCOM⁴ marine monitoring programmes containing over 2000 datasets.

2. THE COFINO MODE

2.1 The Mathematical Basis of the Model

A descriptive overview of the model and examples are given⁵. The details and the justification of the model have been published (Cofino *et al.*, 2000).

Data arise from a measurement process which, when under control, gives an output that can be described by a specific probability density function (pdf). A pdf can be attributed to a particular dataset by adding up the pdfs associated with all the individual measurements (Fig. 2). The overall pdf constructed in this manner is the starting point for the model. Instead of calculating the mean of the data, the model sets out to establish the most probable value, giving the overall pdf. The mathematical procedure borrows the concept of wavefunctions from quantum mechanics.

³Oslo and Paris Commission

⁴Helsinki Commission

⁵Information on the input, statistical output and graphical representations of the Cofino model is given in Annex I

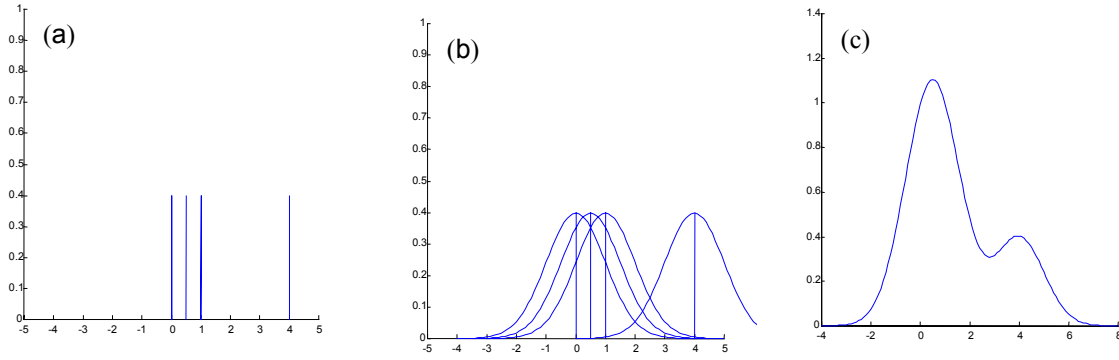


Figure 2. Construction of the overall measurement function of the dataset from the pdfs of the individual measurements. Four individual values (a) from the measurement of a determinant have an uncertainty that can be described by the laboratory measurement function (b). There is considerable overlap for three of these values and a moderate overlap with the fourth value. An overall overlap can be constructed (c) that provides a profile of the overall measurement function.

As an analogue to wavefunctions, observation measurement functions (OMF, ϕ_i)⁶ are defined as the square root of the probability density function which is attributed to the individual observation in question. The set of OMFs forms a space, or a basisset, in which population measurement functions (PMFs²) are constructed. The construction of the PMF Ψ_i is a linear combination of OMF's, i.e. $\Psi_i = \sum c_{ij} \phi_j$. A normalised, squared PMF is a pdf.

In the model, the coefficients are obtained by locating the non-normalised PMF that has the highest probability in the basisset (the maximum in Figure 2). The probability of PMF_i is

obtained as the integral $\int \Psi_i^2 dc$. Mathematically we have to establish the set of coefficients for which the integral $\int \Psi_i^2 dc$ is maximal. The mathematical procedure uses the method of Lagrange multipliers and imposes one additional constraint, that the sum of the squared coefficients is equal to one. The mathematics requires a solution to the eigenvector-eigenvalue equation $Sc = \lambda c$. In this equation, S represents the matrix of overlap integrals.

For example, the matrix element S_{12} is calculated as $\int \phi_1 \phi_2 dc$, i.e. the integral of the product of OMF₁ and OMF₂. S_{12} provides a quantitative measure of how well the two observations agree, taking the respective pdfs into account. It can range between 0 (no overlap) and 1 (100% overlap, the observations have identical pdfs).

The model works in a set of n basisvectors, OMF, to give a total of n eigenvectors c with eigenvalues λ . The eigenvalue λ_i gives the probability in the basisset of the corresponding eigenfunction i . The highest probability and thus maximum value for λ is equal to the number of data n , which is obtained when all data have exactly the same pdf. In this case,

⁶In this and following papers, the terminology is changed from that in [Cofino *et al.*, 2000]. Laboratory measurement function is replaced by observation measurement function, interlaboratory measurement function is now denoted as population measurement function. This modification is applied, as the scope of the model is much broader than interlaboratory studies.

each OMF has the same coefficient that is equal to $\frac{1}{\sqrt{n}}$. The eigenvector with the highest

eigenvalue λ is the PMF ψ_1 is selected and denoted as PMF₁. The remaining n-1 linear combinations are ranked according to probability (i.e. eigenvalue) and are denoted as PMF₂ ... PMF_n. PMF₂ and higher PMFs may sometimes be additional modes, but are frequently only clusters of data ordered according to their degree of overlap. Each squared PMF effectively describes a part of the pdf of the whole data set. When the squared PMFs are summed together over the entire concentration range, the pdf of the entire dataset is reconstructed.

For each PMF Ψ_i the expectation value (\bar{x}) and variance (s_i^2) can be calculated as follows:

$$\bar{x}_i = \frac{\int c * \psi_i^2 dc}{\int \psi_i^2 dc},$$

$$s_i^2 = \frac{\int c^2 * \psi_i^2 dc}{\int \psi_i^2 dc} - \bar{x}^2$$

In addition to the mean and standard deviations of each mode or cluster, the eigenvalues λ enable the quantitative assessment of the degree of comparability and the character (unimodal, bimodal) of the dataset. The program converts the eigenvalue of the mode or cluster proper into a percentage of the overall pdf. The percentage therefore quantitatively describes which fraction of the dataset is accounted for by the PMF in question.

A simple example serves to illustrate the concepts of the model. Three data are reported with means μ_1 , μ_2 and μ_3 and *within-laboratory* standard deviations all equal to s . The associated basisfunctions can be represented by the formula $\varphi_i = \sqrt{N(\mu_i, s)}$. The general form for a PMF is $\Psi_i = c_{1i}\varphi_1 + c_{2i}\varphi_2 + c_{3i}\varphi_3$. The integral $\int \Psi_i^2 dc$ is not normalised and gives the probability of PMF_i in the basisset. This probability depends on the characteristics of the basisfunctions (i.e. means and *within-laboratory* standard deviations) and the coefficients. The values of the probability of PMF_i $\int \Psi_i^2 dc$ can be represented graphically. Two scenarios are used. In scenario 1 the means μ_1, μ_2 and μ_3 are assumed to all be equal. In scenario 2 $\mu_2 = \mu_3$ and μ_1 has a large deviating value so that $S_{12} = S_{13} = 0$. The coefficients $c_1^2 + c_2^2 + c_3^2 = 1$ implies that the value of c_3 is fixed when c_1 and c_2 are chosen. For each scenario the magnitude of $\int \Psi_i^2 dc$ is plotted as function of c_1 and c_2 (Fig. 3).

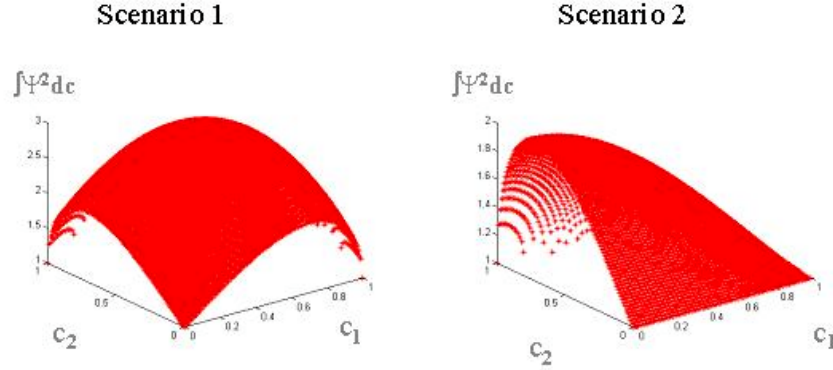


Figure 3. The probability of PMF as function of the coefficients c_1 and c_2 .

In scenario 1 $\int \Psi_i^2 dc$ has a maximum equal to 3 for $c_1=c_2=c_3=1/3\sqrt{3}$. This can be understood as all three basisfunctions are identical – each basisfunction should therefore have the same weight. $\int \Psi_i^2 dc$ is three as it is made up by three pdfs (squared basisfunctions) each having unit area. In scenario 2, a maximum is observed for $c_1=0$, $c_2=c_3=1/2\sqrt{2}$. This maximum has a value 2 as it is determined by two identical pdfs each having unit area. In the model, each datapoint is attributed a pdf with a specific mean and standard deviation. The method of Lagrange multipliers establishes the set of coefficients that give rise to the maximum value of $\int \Psi_i^2 dc$ in the basisset. The eigenvalue belonging to the linear combinations gives the probability of the combination. For scenario 1, the model calculates a mode with mean $\mu_1 (= \mu_2 = \mu_3)$ and eigenvalue 3 (as discussed above the sum of three pdfs each with unit surface area). The eigenvalues (or probabilities) of modes 2 and 3 are zero. All of the probability is contained in the first mode. As the number of observations is 3, the eigenvalue of the first mode is represented as a percentage as $\text{eigenvalue}/(\text{number of observations}) \times 100 = 3/3 \times 100 = 100\%$. The percentages of modes 2 and 3 are both 0%.

For scenario 2, the model calculates for the first mode a mean $\mu_2 = \mu_3$ with eigenvalue 2 (the sum of two pdfs each with unit area, percentage $2/3 \times 100 = 66.7\%$) and for the second mode mean μ_1 and eigenvalue 1 (percentage $1/3 \times 100 = 33.3\%$). The latter corresponds to the observation 1 that does not overlap with 2 or 3 and has unit area.

The model requires estimates of the *within-laboratory* standard deviations in order to define the pdfs and thus the observation measurement functions used in the calculations. In principle, the *within-laboratory* standard deviations could be obtained from the laboratories. This provides the most objective approach and, in most cases, has a higher discriminating power for outliers. It also provides the best insight into the structure of the data. The main disadvantage of using this approach is that it is difficult to readily obtain regular, reliable values for the uncertainty from the laboratories.

As stated before, using the normal distribution approximation (NDA) can circumvent the need to specify the *within-laboratory* standard deviations. The *within-laboratory* standard deviation is constructed using the NDA model.

$$\text{s.d.}_{\text{NDA}} = 1.168 * \text{median}(\text{abs}(x_i - \text{median}(x)))$$

The advantage of the NDA model is that no estimates of the *within-laboratory* uncertainties are needed. However, the disadvantages of the NDA approach are that (i) it is only valid for approximately normally distributed data, (ii) there is less discrimination against a series of outliers and (iii) it gives less insight into the structure of the data.

2.1.1 Left censored values

The model has been extended for use with *left censored values* (LCVs)⁷ by applying the appropriate probability density functions. A straightforward approach can be taken when no assumptions are made regarding the probability density function of LCVs. In such a case, each concentration between zero and the limit of quantification (LOQ) has an equal probability. We can then use the square root of a rectangular probability density function as

basisfunction. Explicitly, when a LCV is reported, the basisfunction is equal to $\sqrt{\frac{1}{\text{LOQ}}}$ in the

interval between zero and LOQ and zero otherwise. These basisfunctions have an expectation value $\int \phi_i^2 dx = \text{LOQ}/2$ and a variance $\int \phi_i^2 c^2 dc - \bar{x}^2 = \text{LOQ}^2/12$. When specific knowledge of the measurement process and the properties of the measured object is available, it would be possible to use other probability density functions. Montville and Voigtman derived pdfs for instrumental limit of detection [2003]. The pdfs can be used when the model is specifically applied to such data. The implicit assumption made with the maximum likelihood method and log probability plotting techniques entails that the LOQs are cut off from the population formed by the numerical values, implying that a concentration just below the LOQ is more likely rather than near zero. To mimic this assumption in a simple way, in this paper a basisfunction has been defined as the square root of a simple triangular

pdf. This triangular pdf has the form $\frac{2}{\text{LOQ}^2} c$ for concentrations between zero and LOQ

and zero otherwise, with an expectation value $\int \phi_i^2 c dc = 2 * \text{LOQ}/3$ and a variance $\int \phi_i^2 c^2 dc - \bar{x}^2 = \text{LOQ}^2/18$.

In the absence of additional information, for applications in this paper and in the QUASIMEME LP studies a rectangular pdf has been assumed. In the application of this model that includes LCVs, the upper limit has been set to LCV \leq mean of the numerical values. Further explanation is provided in Section 6.

2.2 Laboratory Uncertainty for the Cofino Model

The classical way of obtaining the *within-laboratory* uncertainty is to request a replicate number of measurements for each determinand. Normally the minimum number to establish a *within-laboratory* uncertainty would be 6 or more.

This is costly in time and materials for both the LP study provider and each of the participating laboratories. In practical terms such requests are unpopular with the participants and unsustainable in the long term. In addition some laboratories will make the minimum number of measurements while others will make a larger number of replicate measurements and even select data that gives the smallest variance, even though the protocol provided may give the exact instructions. This results in a wider range of *within-laboratory* variances reported. In a recent QUASIMEME exercise for Brominated Flame

⁷Left censored values (LCVs) is the correct nomenclature for “less than” values.

Retardants the *within-laboratory* variance ranges from $\pm 3\%$ to $\pm 42\%$ for one determinand [de Boer & Wells 2002].

An alternative approach would require the participating laboratories to report their uncertainty for each determinand from their Shewhart Control charts and or uncertainty budgets. At the present time clarity over the definitions for uncertainty and therefore the values are needed. Agreed methods of evaluation would be required. Moreover, the information provided would be untraceable and not directly related to the specific exercise and concentration of the determinand. Therefore, this approach is not considered viable.

A model relating the s.d. _{within} to the concentration of the determinand – matrix could be constructed [Gert Asmund – Private Communication 2002]. This can be done retrospectively with LP data. However, this approach requires (i) that the laboratories are under control for the period of establishing the data for the model, and (ii) that the laboratories have participated in the LP studies programme for a number of rounds, ca 10-15 data points would be a minimum. Testing of this method revealed that very precise and accurate data are needed for the algorithm to work. This approach would not be suitable for new participants.

Estimates based on expert judgement could be used for the *within-laboratory* standard deviations from the results of *ad hoc* trials. This approach worked well during the development of the model. However, this is not sufficiently objective and does not relate to the data. Although it maybe a pragmatic solution it is difficult to substantiate with both the participants and the LP studies accreditation bodies.

Although the model using uncertainties supplied by laboratories is a powerful tool to add to the existing techniques of data evaluation, the current ways of obtaining a reliable measure for the *within-laboratory* standard deviation are not readily available at present. Therefore, an alternative approach to the construction of a BandWidth Estimator (BWE) from a selection of the data has been developed as a more sensitive extension to the NDA model.

2.3 BandWidth Estimator for the Cofino Model

The robust statistical methods and, to a much lesser extent, the Cofino NDA model are affected by outliers and the most effective way of dealing with this problem has been to *trim* the dataset to obtain the best estimate of the mean. Within the QUASIMEME programme this trimming has been achieved by removing all data with values $|Z| > 6$ i.e. data more than $6 * \text{target s.d.}$ from the mean. This approach has generally worked well, but trimming is entirely subjective.

However, the NDA model implies that within each dataset, which might be skewed, tailing or perhaps bimodal, there is a set of normally distributed data within the main mode. Therefore by taking a selection or *heart-cut* of the data it should be possible to construct a s.d. _{within} similar to the NDA method for the overall dataset from the selected *heart-cut*. This is the basis of the BandWidth Estimator (BWE), which is a refinement of the NDA approach.

A very similar technique is used for the Kernel Density Estimator (KDE) [Silvermann 1986, Wand and Jones 1995], which has been applied to obtaining a good graphical representation of the distribution of the data for interlaboratory studies [Lowthian and Thompson 2002, AMC 2002]. The Kernel Density method attributes a normal pdf to each data point. All data points are given one and the same standard deviation, obtained from the KDE, h , which is derived from the interquartile range (IQR). The overall data distribution graph is obtained by summing the individual pdfs.

The KDE thus also takes a *heart-cut* to establish the *within-laboratory* standard deviation to be used to construct the graphs. This approach also reduces or eliminates the effects of data that may not belong to the population.

In order to develop the BWE approach, two questions require answers:

1. What formula should be used to the construct s.d. _{within-laboratory} from the *heart-cut*?
2. What criteria should be used to obtain the *heart-cut* of a dataset?

In the course of development of the BWE, two guiding principles were maintained: (i) the BWE should produce the mean and standard deviation of normally distributed datasets and (ii) the BWE should be as robust as possible to '*outlying*' data. i.e. Outlying data may be defined as values which do not belong to the population which make up the most probable value of the dataset.

2.3.1 The construction of s.d. _{within-laboratory} from the *Heart-cut*

The Cofino NDA model provides a value for $\delta_{\text{within-laboratory}}$ on the basis of all data

$$\delta_{\text{NDA}} = 1.1682 * \text{MAD}$$

δ_{NDA} is a construct and MAD is the median absolute deviation. The basis for this formula is in finding the mean and standard deviation of a true normal distribution when normal distributions are used as basisfunctions and when all data are given δ_{NDA} as *within-laboratory* standard deviation. True normal distributions were used to see whether a similar formula could be derived for certain ranges within the distribution, e.g. the IQR or the range between the 40% and 60% percentiles. This appeared to be the case. The relationship $\delta_{\text{bwe}} = \alpha * \text{MAD}$ could be derived with α depending on the fraction of data within the normal distribution used. Thus, when all data are used, $\alpha = 1.1682$ and the formula's for δ_{bwe} and δ_{NDA} become identical. The relationship between the fraction of the data used and the value of α is given in Figure 4. For truly normal distributions, δ_{bwe} and δ_{NDA} should have the same

magnitude regardless of the fraction of data taken. In practise, differences occur owing to the distribution of data. In addition, as the δ_{bwe} is based on only a subset of a data, the uncertainty in its value is higher.

2.3.2 Criteria for the BWE *Heart-cut*

For a normal distribution the mean would occur approximately at the 50 percentile. However, for a dataset that is contaminated with outliers, the mean may be displaced from the 50 percentile, depending on the nature of the skewed distribution. It is assumed that the first mode of the dataset is made up by results from ‘*well behaved*’ laboratories and has an underlying normal distribution. When the highest mode is not found at the 50 percentile it would be inappropriate to base the BWE on a straightforward IQR, as it would capture an asymmetric fraction of an assumed normal distribution. In our approach we therefore first calculate the position of the first mode of the dataset using the NDA model and position the *heart-cut* symmetrically around the position of this mode.

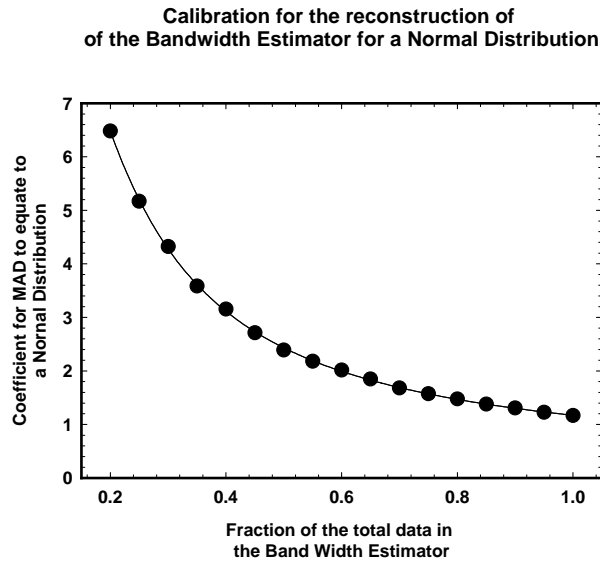


Figure 4. The factor α was obtained by constructing a normal distribution of 1000 data points. The Cofino NDA model was used to establish the population characteristics of the data. The data was increasingly trimmed leaving a *heart-cut* about the Cofino mean and the Cofino NDA model used again to establish the population characteristics. The factor α was calculated and a curve fitted with an inverse 7th order fit.

In descriptive terms, the BWE is based on a *heart-cut* about the main mode of the data (Fig. 5). The δ_{bwe} is constructed from data contained in this *heart-cut*.

The calculation for the BWE is carried out as follows. Firstly, the expectation value of the full dataset is calculated using the Cofino NDA model. The *heart-cut* is defined as:

$$\text{mean}_{PMF1} \pm 2 * \text{s.d.}_{PMF1}$$

This *heart-cut* is assumed to contain the normally distributed data that make up the first mode. The NDA model can be used with these data. It is possible, however, that the *heart-cut* still contains data that do not belong to the population. Furthermore, the estimation of s.d._{NDA} (and/or s.d._{bwe}) may be affected by the distribution of data within the *heart-cut*. Therefore, the range of the selected data is reduced incrementally in steps of 5%. An estimate of s.d._{bwe} is obtained for each fraction. In this manner, a series of values for s.d._{bwe}

is produced. The stepwise reduction of the fraction is continued until the fraction of data is 30% or until the number of observations (NObs) is less than 11. The Cofino NDA model is then used to calculate the mean $s.d_{bwe}$ of this series. The mean $s.d_{bwe}$ is used as the BWE representing the core of laboratories that are in good agreement.

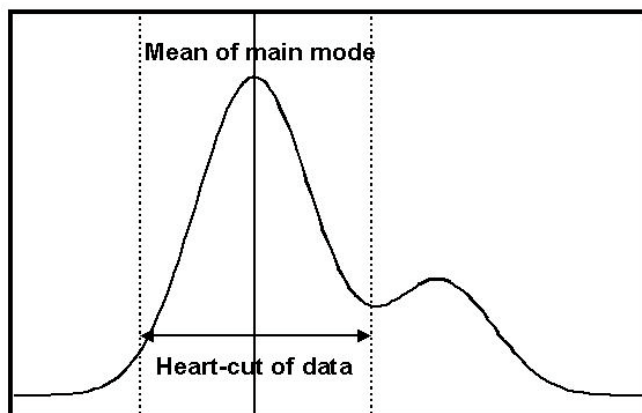


Figure 5. A distribution profile of a bimodal dataset. The estimation of the mean and standard deviation of the main mode of the data can be obtained, in situ, without removal of outlier or separating the modes using the Cofino Bandwidth Estimator (BWE) model. The BWE is obtained by first finding the mean of the main mode using the Normal Distribution Approximation (NDA) and then by taking a *heart-cut* of ± 2 s.d. This revealed data is used to establish the BWE.

The Cofino model, used in this way, provides the BWE estimators of the mean and s.d. of the entire dataset. The limitation of 30% and NObs = 11 are empirical boundaries that have been selected to obtain the BWE using the maximum amount of any dataset. These boundaries are not sensitive to the outcome of the model. The schematic flow diagram of the calculation process is given in Figure 6. The performance of the BWE method and the NDA method for normal distributions is given in Table 1. In this table, the results are also compared with those obtained with robust statistics and straightforward calculation with conventional statistics. Table 1 shows that both BWE and NDA approach perform well on normally distributed datasets.

The BWE is based about the main mode of data that are in good agreement. This value is quantitative, traceable to the dataset and the percentage of the data used for the BWE is known. This BWE provides a better estimate of $s.d_{within-laboratory}$ as it selects the data in good agreement from within a complex dataset. As a result $s.d_{within-laboratory}$ is smaller than $s.d_{NDA}$ and provides more information on the structure of the data and a better estimate of the population characteristics on ALL data including tailing values and outliers.

Although either of the Cofino models can be used to evaluate data from LP studies there are some basic difference between them. The NDA approach assumes an underlying normal distribution for the entire dataset and uses all data to calculate the $s.d_{NDA}$. The BWE approach assumes that within the dataset, the mode ± 2 s.d. will represent normally distributed data provided by the best performing laboratories. These data define the *heart-cut* that is used to calculate $s.d_{bwe}$. For normally distributed data $s.d_{bwe}$ and $s.d_{NDA}$ are very similar. In our opinion this provides a better estimate for the population characteristics when the overall distribution deviates significantly from normality. An example of this can be seen for chlorobiphenyl CB 28 in biological tissue (Fig. 7). The chromatographic separation of CB 28 from the congener CB 31 is very dependant upon optimum chromatographic conditions of flow rate, carrier gas, column selection, programming, the overall condition of

the capillary column, and the efficacy of the sample preparation. Any deterioration will reduce the separation and the apparent concentration is a sum of CB 28 and the adjacent CB 31.

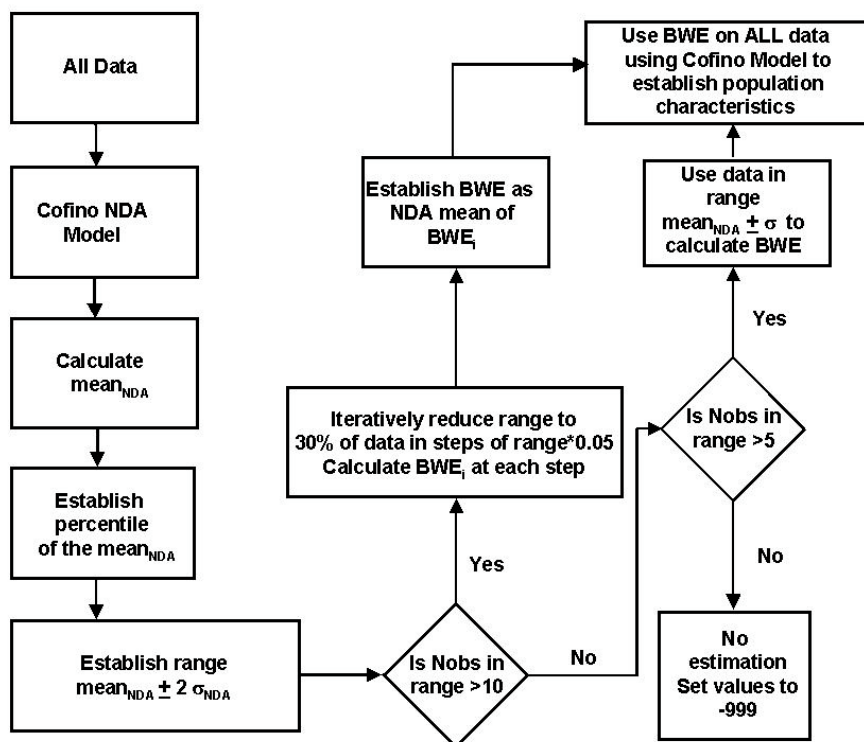


Figure 6. Schematic diagram of the method for calculating the Cofino model BandWidth Estimator.

For these data, Figure 7 shows δ_{NDA} is higher compared with the δ_{bwe} based on a *heart-cut* of the data (Fig. 7). The BWE provides more structural information of the population density functions shown in the centre two traces and also of the Kilt plot⁸ (Overlap matrix). The individual modes are substantially more distinct, and values and specific measurements are more definitively associated with these modes. In many cases, the BWE, based on the *heart-cut* of data brings more information on the structure of the data and consequently the assessment.

The main mode associated with a large percentage of the data reflects the measurement of CB 28, with a satellite peak with fewer data which is more likely to be the measurement of unresolved CB 28 and CB 31. Since the magnitude of CB 28 and CB 31 can often be quite similar a value for PMF₁ of 0.2 µg / kg for CB 28 and a value of ca 0.4 µg / kg for CB 28 + CB 31 is in keeping with the type of result which may be expected.

⁸Kilt plot is the graphical representation of the overlap matrix – see Annex I.

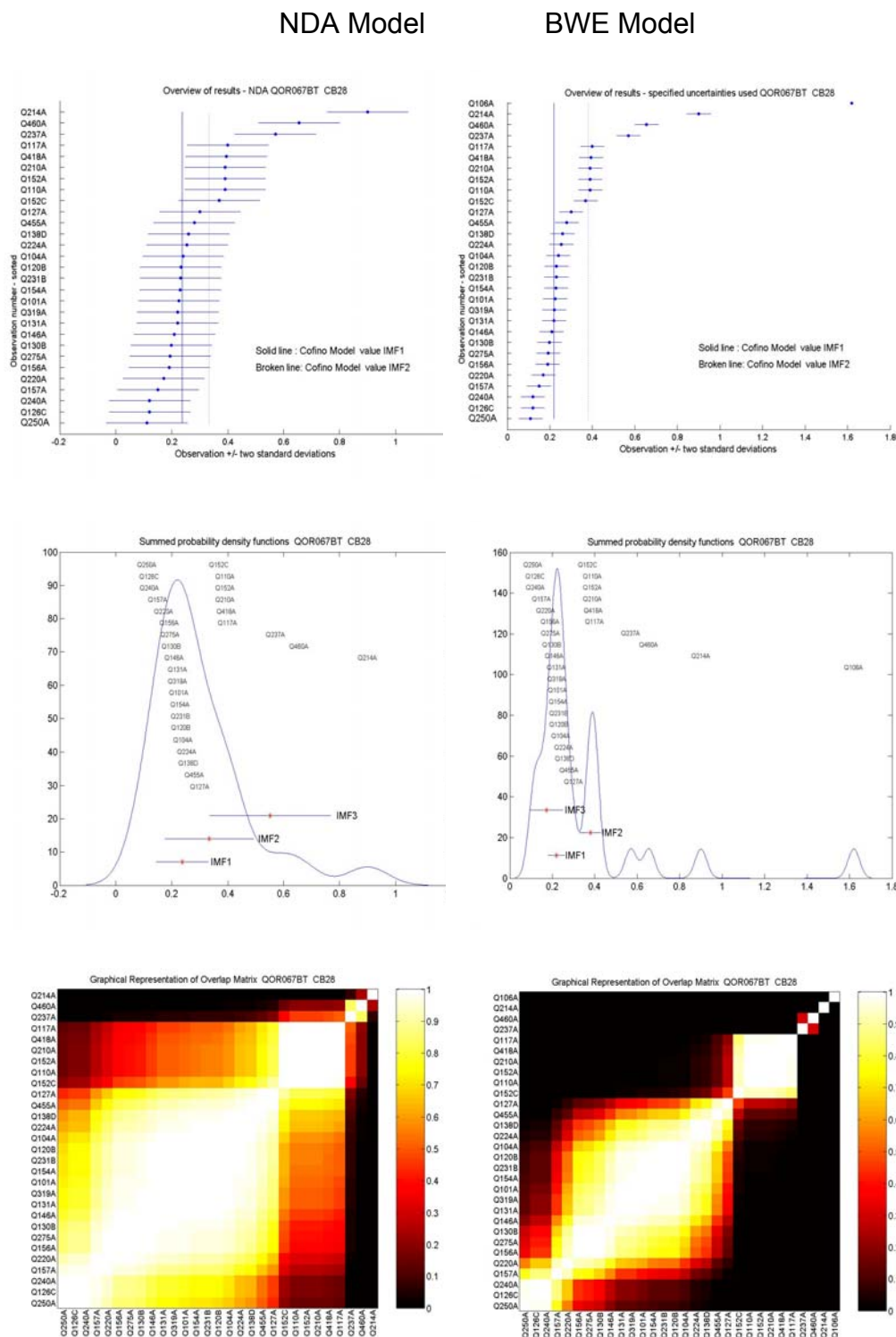


Figure 7. Chlorobiphenyl CB 28 in biological tissue. A comparison of the different structural information obtained with the NDA and the BWE models. The three graphical outputs of the ranked overview of the data (upper graphs), the summed pdfs (centre graphs) and the Kilt plot (Overlap matrix) in the lower graphs. Greater structural detail is obtained with the BWE model showing the bi / multimodal nature of the distribution.

3. VALIDATION OF THE COFINO MODEL WITH TEST DATA

Information on the input, statistical output and graphical representations of the Cofino model is given in Annex I.

3.1 Normally Distributed Test Data

The BWE is an extension of the normal distribution assumption (NDA) in which an underlying normality of the data or a selection of the data is implied. This hypothesis maybe tested against a normally distributed dataset, where the results should be comparable, regardless of the method of calculation.

A series of normally distributed data ($N(\mu, s)$) were randomly generated, where $s = 2, 5, 10$ and 30 and $\mu = 100$. Each of the series was generated using the MATLAB randn command for 200, 100, 50, 20 and 10 data.

Each dataset was evaluated using normal and robust statistics [AMC 1989a and b], as well as the Cofino NDA and BWE models. The results are given in Table 1.

Each separate series of normal distributions were randomly generated x1000 and sorted. In this way a matrix with NObs⁹ (200,100,50,20,10) columns and 1000 rows was created. The mean was obtained for each column creating a vector with NObs_{column}. This vector provided a smoothed normal distribution against which to compare the output of the models. The results for the normal arithmetic mean, the Cofino NDA and BWE model mean, and robust mean can be compared in Table 1. The percentage of data included in the first mode of the NDA and BWE models are also given. The results for the arithmetic s.d., the Cofino NDA and BWE model s.d., and robust s.d. are also given in Table 1. The percentage of NObs in the BWE and the percentile of the mean are also given for each normal distribution tested.

This method of smoothing the randomly generated normal distribution has produced a near perfect series of normally distributed data against which to evaluate the calibration of the models.

The mean values range from 99.55 to 100.08 with a maximum difference of <0.2% between any model and the arithmetic mean. The expected value in all cases was 100.00.

The s.d. for each of the series of laboratory numbers was selected to cover the range experienced in many of the performance studies.

The difference between the arithmetic s.d. and the Cofino NDA model s.d. varied from 0.36 to 5.08%, with an average of 2.08%. The greater difference occurred with the smaller number of laboratories.

The differences for the s.d. for the BWE ranged from -2.9 to 5.7% with an overall value of 0.26%, showing good agreement with the arithmetic s.d.

The robust s.d. shows a wider difference with the arithmetic s.d. ranging from 0.75 to 13.7 with an overall value of 4.85%.

This series of calibrations gives a clear indication of the high level of comparable summary statistics for the 3 models, Cofino NDA and BWE and the robust (RSC-AMC) model. The mean values for all models give values very close to the arithmetic mean.

⁹ NObs = Number of Observations.

In each case, as might be expected, the highest differences in the s.d.'s between models occur with the datasets with the smallest number of observations. A difference of only 5% with just 10 observations, and, 2.5% for higher number of values, demonstrates that the BWE has been correctly calibrated for an underlying normal distribution.

3.2 Comparison with Reference Data

Two sets of reference data from the National Institute of Standards and Technology [NIST] were obtained. Both sets of data were taken from the NIST univariant data bank. The first set, 'Michelson' [Dorsey, 1944] was based on the measurement of the velocity of light with a mean value of 299.8524 m/s and s.d. ($n - 1$) of 0.079 m/s (NObs = 100). The full value of the s.d. was reported to 14 significant figures.

The second set of data, 'Mavrio' [Mavrodineanu, 1971] was based on the transmittance value for a filter with a certified mean of 2.001856 and s.d. ($n - 1$) of 0.00429 (with 15 significant figures).

The summary statistics obtained with the normal and robust statistics and the Cofino NDA and BWE model are given in Table 2 and the histograms are given in Figure 8 for Michelson and for Mavro.

TABLE 1

Comparison of mean and standard deviation of a randomly generated normal distribution using normal statistics, robust statistics and the Cofino model with normal distribution assumptions (NDA) and the bandwidth estimator (BWE).

Sample	Expected Mean	Arithmetic Mean	NDA Mean	BWE Mean	Robust Mean	NDA % Data	BWE % Data
200Labs	100	100.00	100.00	100.00	100.00	76.51	75.18
200Labs	100	100.00	99.99	99.99	100.00	76.55	75.23
200Labs	100	99.95	99.96	99.96	99.96	76.49	75.02
200Labs	100	99.98	99.97	99.97	99.98	76.50	75.16
100Labs	100	100.00	100.01	100.01	100.01	76.42	75.08
100Labs	100	100.00	100.00	100.00	100.00	76.56	75.25
100Labs	100	100.01	100.03	100.03	100.02	76.66	75.22
100Labs	100	100.02	99.99	99.98	100.01	76.43	74.99
50Labs	100	100.02	100.03	100.03	100.02	76.70	75.23
50Labs	100	100.01	99.99	99.99	100.00	76.56	75.19
50Labs	100	99.96	99.99	99.99	99.97	76.85	75.47
50Labs	100	100.08	100.01	100.01	100.06	76.81	75.54
20Labs	100	99.99	99.99	99.99	99.99	77.34	77.43
20Labs	100	100.06	100.05	100.05	100.05	77.29	77.25
20Labs	100	100.03	99.97	99.97	100.02	77.25	77.24
20Labs	100	99.86	99.80	99.80	99.86	77.06	76.95
10Labs	100	99.99	99.99	99.99	99.99	77.67	77.79
10Labs	100	100.04	100.02	100.02	100.04	78.16	78.48
10Labs	100	99.96	99.86	99.86	99.96	77.63	77.73
10Labs	100	99.72	99.55	99.56	99.72	78.19	78.35

TABLE 1 (CONT)

Comparison of mean and standard deviation of a randomly generated normal distribution using normal statistics, robust statistics and the Cofino model with normal distribution assumptions (NDA) and the bandwidth estimator (BWE).

Sample	Expected s.d.	Normal s.d.	NDA s.d.	BWE s.d.	Robust s.d.	% NObs in BWE	Percentile of Mean
200Labs	2	1.99	2.00	1.94	2.00	72	50
200Labs	5	4.98	5.02	4.87	5.03	78	50
200Labs	10	9.92	9.98	9.64	10.00	80	50
200Labs	20	19.88	19.98	19.37	20.03	68	50
100Labs	2	1.98	1.99	1.93	2.01	74	50
100Labs	5	4.96	5.01	4.86	5.03	74	50
100Labs	10	9.93	10.05	9.72	10.08	63	50
100Labs	20	19.82	19.89	19.24	20.06	73	50
50Labs	2	1.97	1.99	1.93	2.02	70	50
50Labs	5	4.92	4.98	4.83	5.05	70	50
50Labs	10	9.80	10.00	9.69	10.07	76	50
50Labs	20	19.54	19.96	19.38	20.14	72	50
20Labs	2	1.93	1.99	2.00	2.04	70	50
20Labs	5	4.80	4.97	4.96	5.10	65	50
20Labs	10	9.60	9.90	9.90	10.16	65	50
20Labs	20	19.22	19.73	19.68	20.34	65	50
10Labs	2	1.87	1.93	1.94	2.12	60	50
10Labs	5	4.65	4.87	4.91	5.27	60	50
10Labs	10	9.41	9.72	9.74	10.67	60	50
10Labs	20	18.79	19.74	19.81	21.30	60	50

TABLE 2

Comparison of mean and standard deviation of NIST reference data (Mavro and Michelso) using normal statistics, robust statistics and the Cofino NDA and BWE model.

Sample	NIST Certified Value	Arithmetic Mean	NDA Mean	BWE Mean	Robust Mean	NDA % Data	BWE % Data
Mavro	2.00186	2.001856	2.001767	2.001694	2.001845	77.07	65.10
Michelso	299.852	299.8524	299.8494	299.848	299.8527	72.10	66.30
Sample	NIST Certified s.d.	Arithmetic s.d.	NDA s.d.	BWE s.d.	Robust s.d.	% Data in BWE	Percentile of Mean
Mavro	0.000429	0.000429	0.000445	0.000327	0.000463	84	50
Michelso	0.079	0.079	0.069	0.061	0.079	76	49

The Kolmogorov – Smirnov (K-S) test for normality was applied to both of these sets of data. The Michelso data passed the test with a K-S distance of 0.083, $p=0.083$ while the Mavro data failed, K-S distance of 0.143, $p=0.011$. Inspecting the distribution of the Mavro data there is clear evidence of bimodality in the data, which accounts for the slightly lower Cofino BWE mean of 2.00169 compared with 2.00186 for the arithmetic mean. The arithmetic mean of Mavro data, less the trimmed values of one of the modes, as indicated in Figure 8 is 2.00168 which is very comparable to the BWE mean.

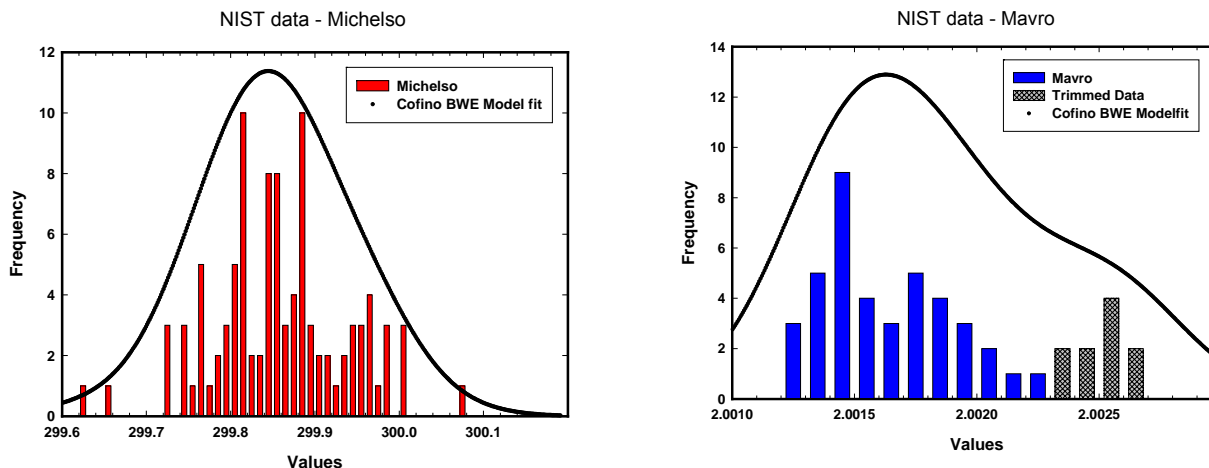


Figure 8. Histogram and the Cofino BWE Model fit of the NIST reference data 'Michelso' and 'Mavro'. The 'Michelso' data passes the Komologrov-Smirnov test for normality, but the 'Mavro' data fails this test.

3.3 Detection of Bimodality

3.3.1 Sensitivity to separation

One aspect of the power of the Cofino model is that it can provide clear information on the distribution of the data and on the presence of bi or multi-modality, which may subsequently be attributed to specific causes, such as differences in performance of the analytical method.

When assessing the structure of a dataset it is important to know how reliable the Cofino BWE model is for detecting truly bimodal data. Over-sensitivity may result in artefacts in the structure while insensitivity would leave differences in the distribution of the data undetected.

A series of normally distributed datasets were randomly generated using the MATLAB randn command to give $N(100,2)$ with 50 data and $N(100+\hat{\epsilon},2)$ with 30 data, where $\hat{\epsilon}=1:10$.

The two sets of data for each degree of separation $N(100,2) + N(100+\hat{\epsilon},2)$ were combined to provide bimodal dataset with a separation of $\hat{\epsilon}$ between the modes. Each combined dataset was evaluated using the Cofino BWE model. The datasets were also evaluated using the Kernel Density Estimator (KDE) [AMC_a 2001] to provide an additional comparison with the overall probability measurement function profile.

The graphical results are given for a 2,4,6 and 8% separation ($\hat{\epsilon}=2,4,6$ and 8) for the pdfs and the KDE profile overlaid together, and for the Kilt plot in Figure 9.

At $\hat{\epsilon}=2\%$ there is an elongation of the area of overlap (white) in the Kilt plot and a peak and distinct shoulder on the KDE profile. The profile of the PMF generated by the Cofino BWE model PMF is broad, but there is no distinct evidence of bimodality.

At $\hat{\epsilon}=4\%$ there is clear evidence of two separating (white) areas of common overlap in the Kilt plot and two distinct peaks with the KD profile. The pdf profile from the BWE now shows an asymmetric shoulder, but still no obvious evidence of bimodality as seen in the Kilt plot.

However, with $\hat{\epsilon}=6\%$ there is very clear evidence of bimodality in all 3 plots with the KDE and pdf profiles being very similar. The overlap matrix shows four very distinct regions with the two modes with a high degree of overlap within the mode and very little between mode interaction. Beyond this at $\hat{\epsilon} = 8\%$ the two modes are completely distinct and very clearly evident with all profiles.

These data set the threshold to detect bimodality at between 4% and 6%. The Kilt plot is the more sensitive to this separation and provides a clear map of the overlap interactions of all contributing values. The KDE profile provides good resolution and detected bimodality at a lower degree of separation compared with the pdf using the values of h_{opt} and the auto BWE. Both the KDE and the BWE can be set to a higher sensitivity, but this is most likely to falsely trigger the detection of two modes where they most likely do not exist.

The KDE provided a good graphical comparison, but no quantitative information. The Cofino BWE model gives the summary statistics for each mode. The expectation value of each mode provides an estimate of the mean of each of the separate datasets.

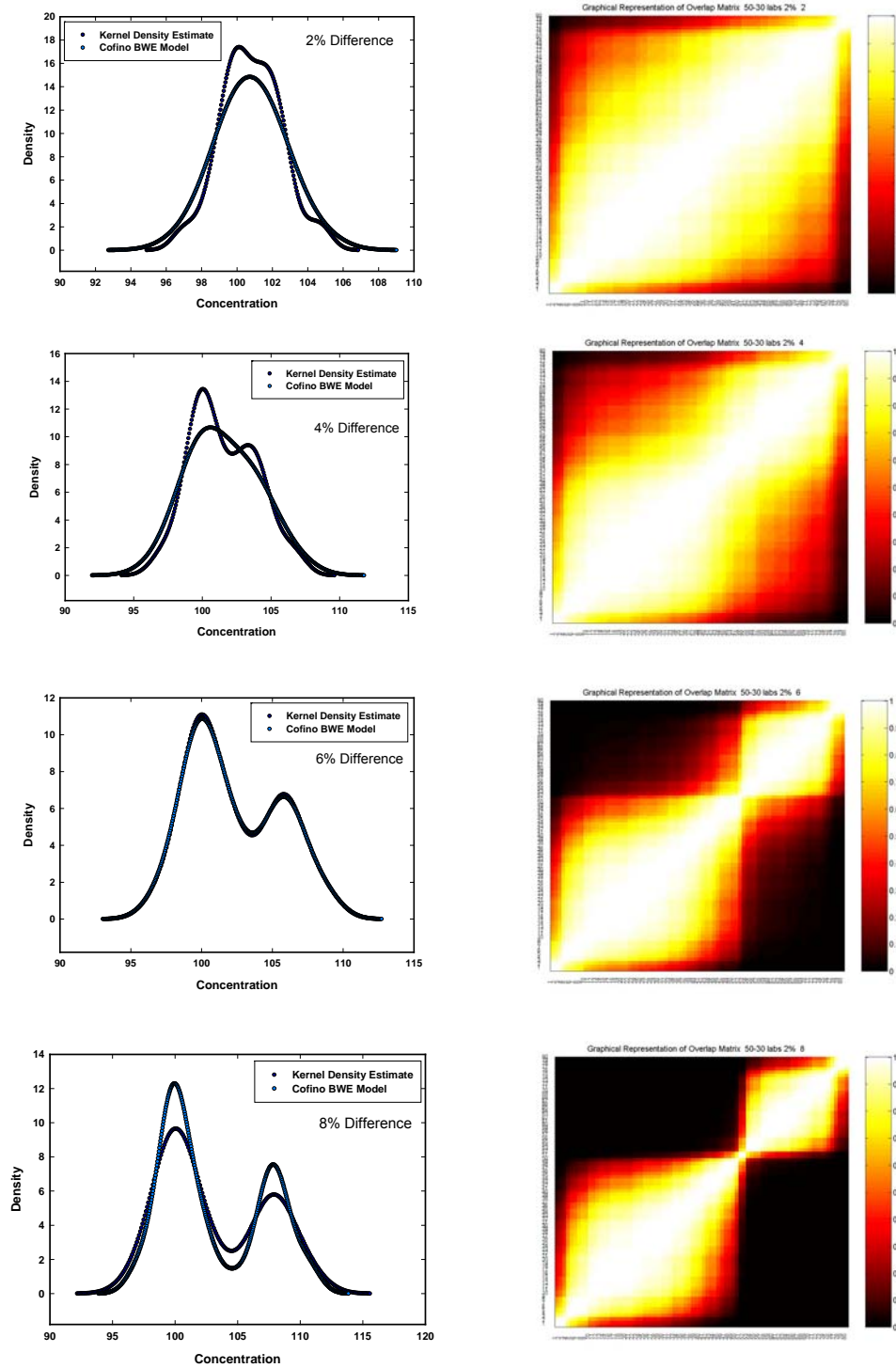


Figure 9. The population density functions and the Kilt plot (Overlap matrix), plots for two randomly generated normal distributions, $n_1 = 50$, $n_2 = 30$. Both distributions have a s.d. of ± 2 and are separated by means differing by 2,4,6 and 8%. The pdfs are generated by the Cofino BWE model and the Kernel Density Estimator [AMC Technical Brief No 4 2001].

In Figure 10 the percentage separation of the two modes is plotted against (i) the expectation value of the 1st and 2nd mode, (ii) the ratio of the % data associated with PMF₁ and PMF₂ and (iii) Student's t value. This provides a clear indication of how well the two modes are separated. Between 4% and 6% little separation is detected and a strong influence on the expectation values of both modes by each set of data is evident. As the separation is increased, the expectation value of PMF₁ falls close to the independent value (100), while the expectation value of PMF₂ rises to its separate, independent value. At the same time the ratio of % of data in the two modes falls to that of the two independent datasets $50/30 = 1.666$.

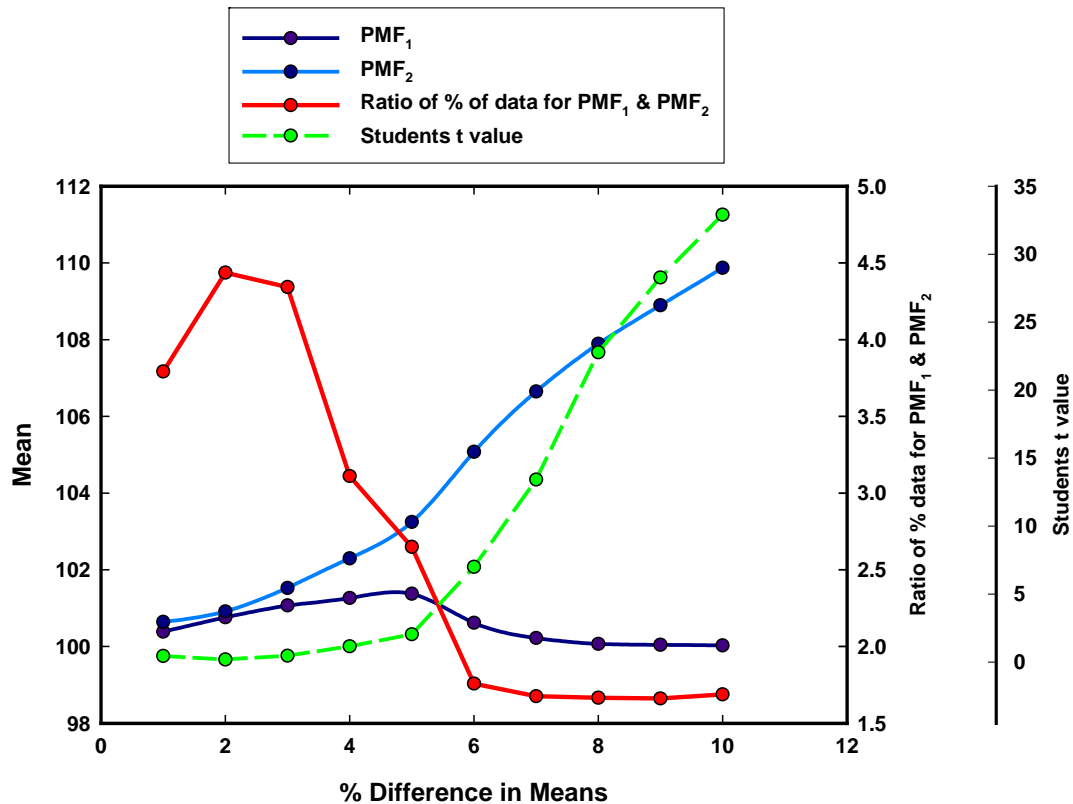


Figure 10. The ability of the Cofino BWE model to identify the separate modes in a bimodal distribution is dependent on the difference between the means and the variance of the data. Data in the two modes ($\text{NObs}_1 = 50$, $\text{NObs}_2 = 30$, $\text{s.d.} = 2$, $\text{Mean}_1 = 100$) are separated where there is a 5% difference in the means.

The identification of bimodality is primarily dependent on two factors. The first, as described above, is the degree of separation between the two modes (Fig. 10). The Student's t critical value occurs between 5-6% separation, after which the two modes are significantly separate.

3.3.2 Percentage of data in each mode

The second key factor to identifying separate modes in data is the relative NObs in each mode. As the separation between the two modes increases the mode with the greater percentage of data will become the primary PMF.

In the example given here the mode associated with PMF₁ contained 62.5% of the data associated with the mode for PMF₂, containing 37.5% (50:30). However, when the balance

of data in the two modes becomes closer to 50:50 it is more difficult to identify and isolate the two modes.

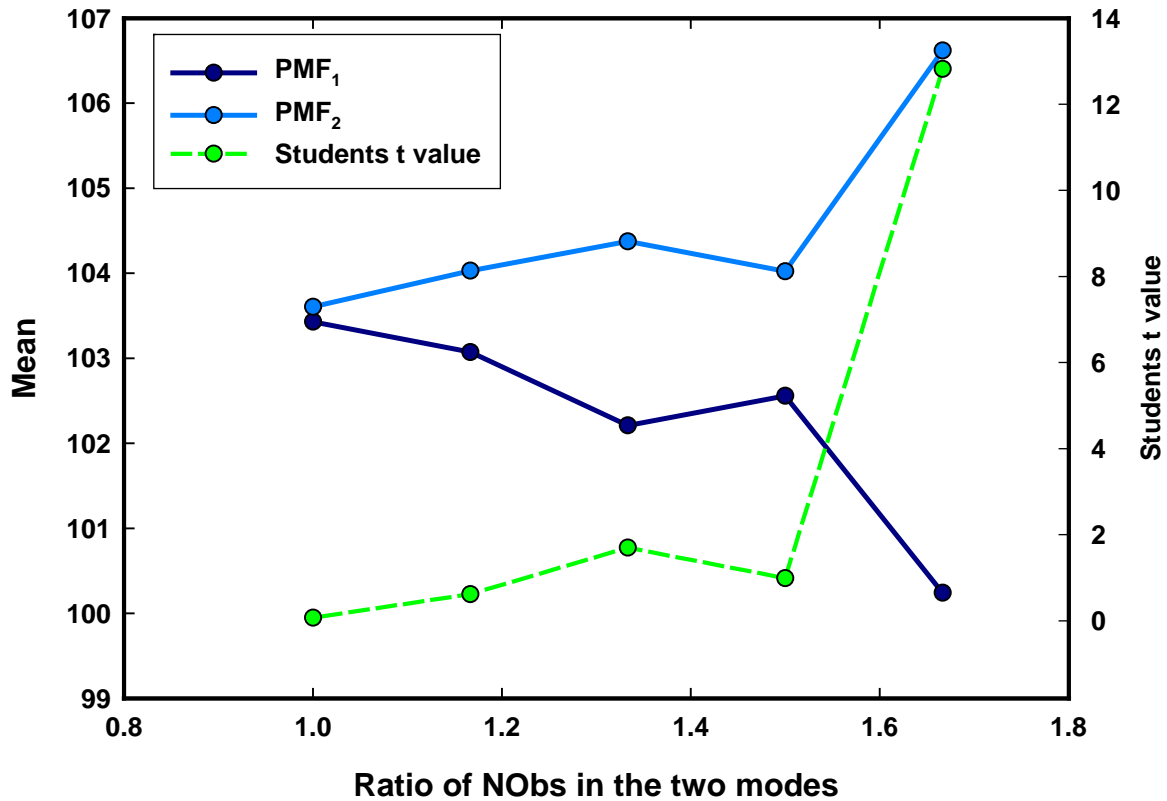


Figure 11. The ability of the Cofino BWE model to identify the separate modes of a bimodal distribution is dependent on the percentage of data in each mode. Data in the two modes have a mean_{mode 1} = 100 and a mean_{mode 2} = 107, s.d. = 2. The ratio of the data must differ by > 1:1.5 to allow the separation and characterisation of the modes.

A series of normal distributions were randomly generated, combined and analysed using the Cofino model: one series with N (100,2) and with NObs from 50 to 30 in increments of 5; a second distribution with N (107,2) and NObs = 30. The values for the means associated with PMF₁, and PMF₂ are plotted against the ratio of NObs in the two modes. Superimposed on this plot is the value of the Student's t value for each of the sets of data (Fig. 11).

With a ratio of 1.66 NObs for the two datasets at 50 and 30 there is a clear separation. Both expectation values for the data associated with PMF₁ and for PMF₂ are close to their independent values 100.2 (100) and 106.6 (107) respectively. However, by decreasing the ratio of NObs to 45:30 the separation between the modes begins to collapse with the mode associated with PMF₁ at 102.6 and PMF₂ at 104. The Student's t value is less than $t_{critical}$ and the PMF profiles show no separation.

Unlike other techniques, the Kilt plot continues to identify the clear distinction in the data, but because there is no predominant mode the overall effect is for each of the expectation values of PMF₂ to adopt similar values. This occurs primarily when adjacent modes have a similar number of observations associated with them. In such a circumstance the data in the two modes can be identified by the overlap matrix (Fig. 12) and each can be treated separately. For all other eventualities most datasets can be assessed in a single calculation.

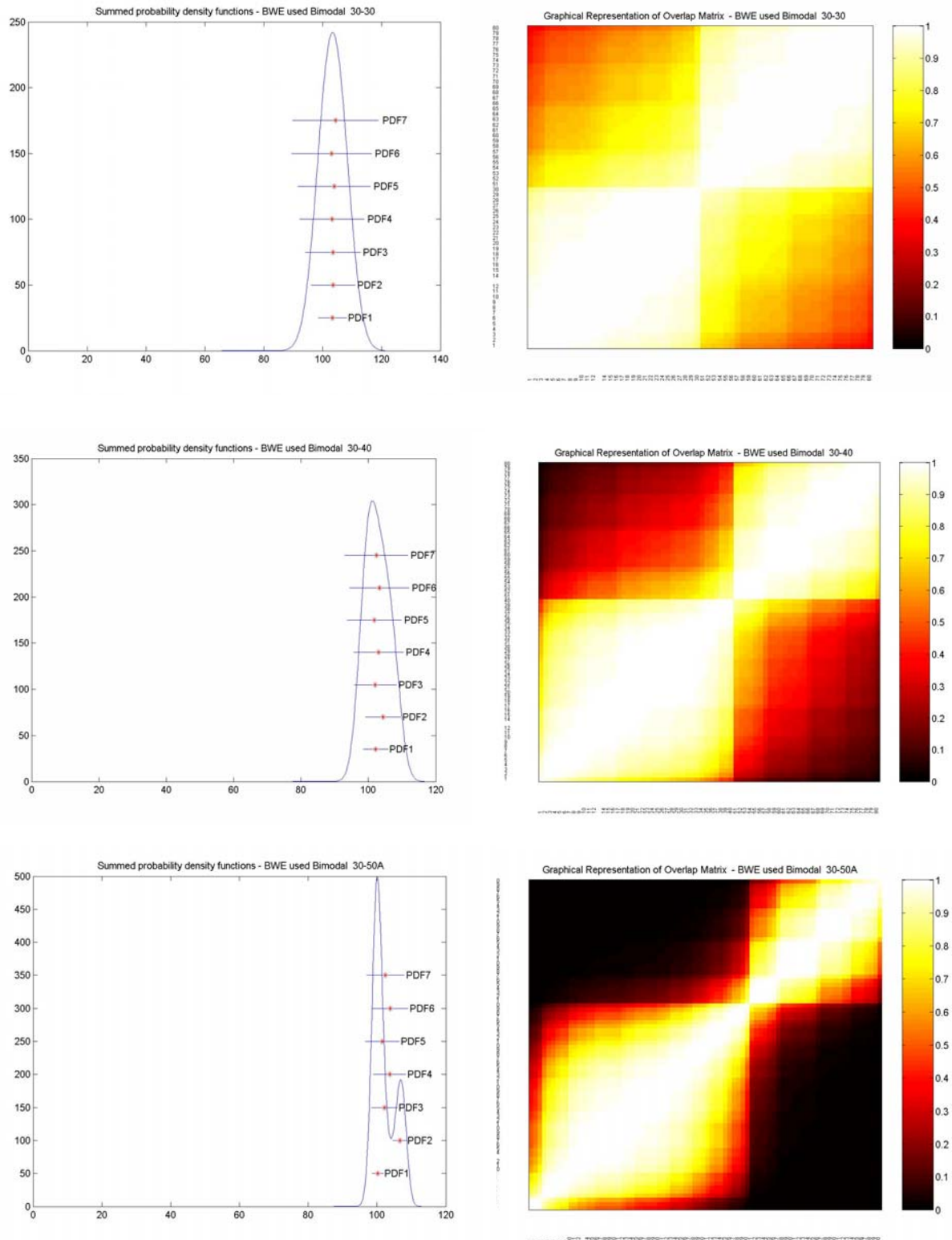


Figure 12. The Kilt (overlap matrix) plots and summed pdfs for a bimodal distribution with modes of mean 100 and 107, s.d.=2. The bimodality is only distinct with the summed pdfs when the ratio of the NObs is $> 1:1.5$, e.g. NObs 30:50. At NObs 50:50 the distinction remains clear on the Kilt plot.

The Cofino Model: A Handbook to Evaluate Laboratory Performance Studies

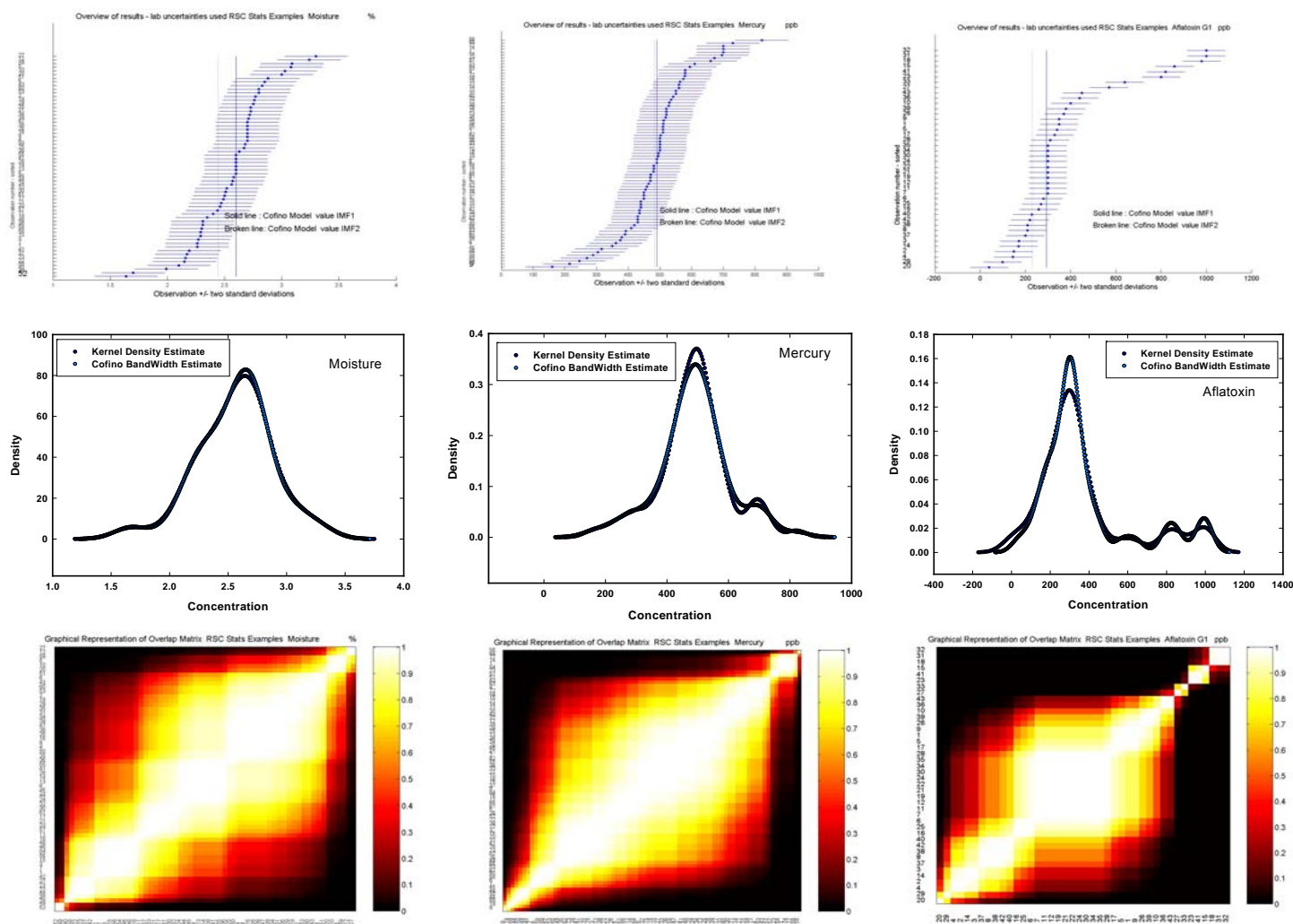


Figure 13. Evaluation of six data sets taken from Lothian and Thompson (2002) using the optimum Kernel Density Estimate and the Cofino BWE Model. (i) Moisture (n=61), (ii) Mercury (n=77), (iii) Aflatoxin (n=92),

The Cofino Model: A Handbook to Evaluate Laboratory Performance Studies

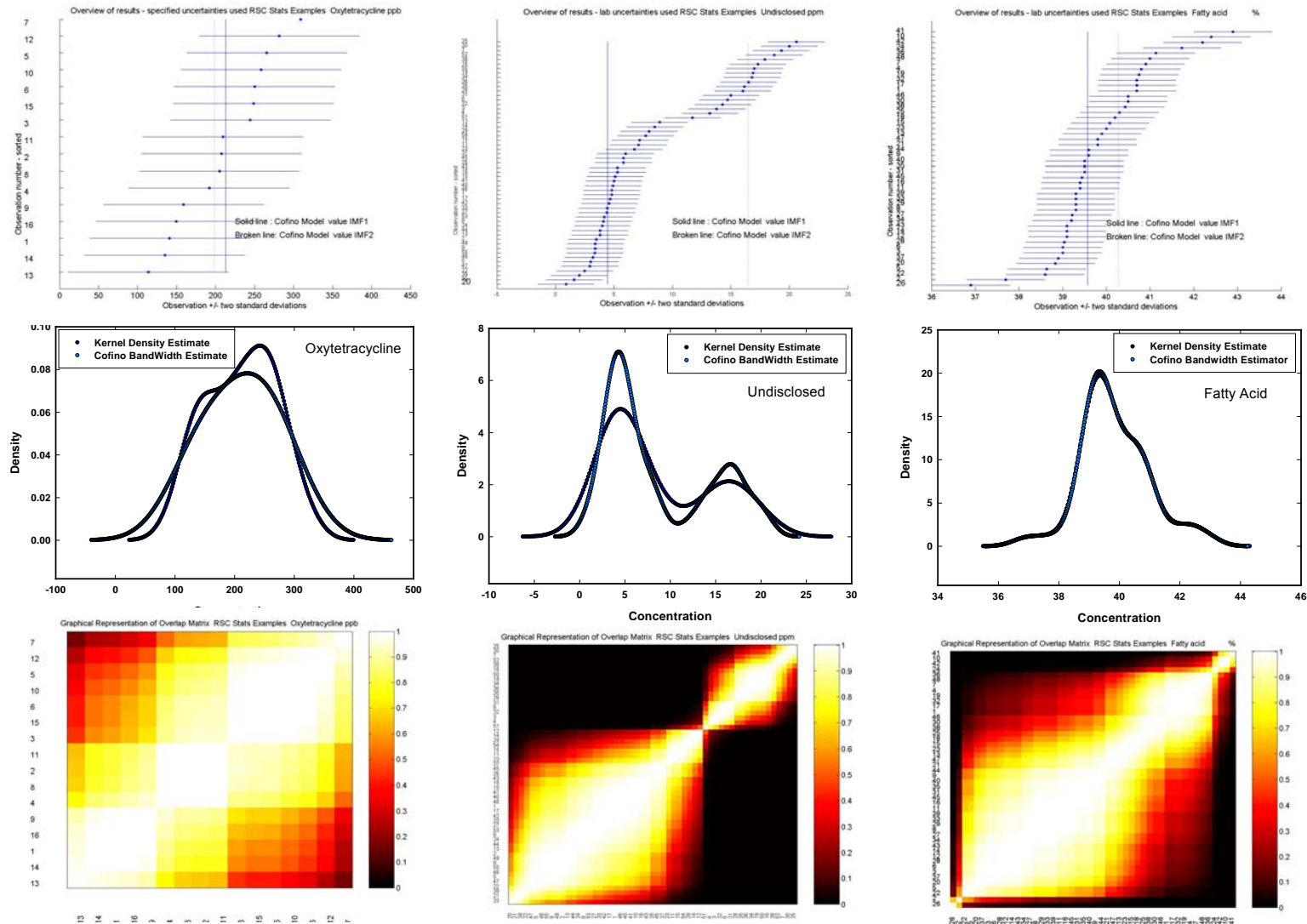


Figure 13 (cont).

Evaluation of six data sets taken from Lothian and Thompson (2002) using the optimum Kernel Density Estimate and the Cofino BWE (iv) Oxytetracycline (n=16), Undisclosed determinand (n=55), and fatty acid (n=48).

4. COMPARISON OF BWE AND KERNEL DENSITY ESTIMATOR FOR DIFFICULT DATASETS

The Cofino model [Cofino *et al.*, 2000] and the application of the KDE [Lothian and Thompson, 2002] have been developed completely independently and are based on two different concepts, yet the pdf output is very similar. The independent philosophy and development of the two methods have permitted an invaluable level of mutual validation.

The data published in support of the KDE to provide a profile of data from proficiency tests was used in conjunction with the Cofino BWE model for six datasets [Lothian and Thompson 2002]. An AMC Excel *add-in* was available to generate the data distribution (pdf) profiles using the KDE.

The ranked means ± 2 s.d. and the overlap matrix for each of the six datasets are shown in Figure 13, along with the comparison of the summed population density function (pdfs) and the Kernel Density pdf.

In each case the Kernel Density profile was generated using h_{opt} given by:

$$0.9 * \min(\text{s.d.}, \text{IQR}/1.34) * n^{-0.2}$$

Where IQR is the inter-quartile range and n is the number of observations [Silverman 1986, AMC 2001a]. The level of agreement between the Cofino BWE summed pdfs and the KDE output are in very good agreement for each of the 6 datasets, although for the KDE approach the graphs are primarily qualitative (Table 3).

TABLE 3

A Comparison of the Bootstrap and Robust Estimation with the Cofino NDA and BWE model for difficult data.

Determinand	NObs	AMC ¹⁰	AMC	Cofino model	
		Estimate	Method	NDA	BWE
Moisture	61	2.29 (0.03) 2.67 (0.04) 3.06 (0.5)	Bootstrap-trimmed Bootstrap-trimmed Bootstrap-trimmed	2.6 (0.032)	2.6 (0.026)
Mercury	77	489.0 (10.6)	Robust Means(RM ¹¹)	489.7 (9.4)	490.3 (6.5)
Aflatoxin	42	328 (25) 283 (17)	RM RM-trimmed data	285.7 (17.2)	292.6 (9.4)
Oxytetracycline	16			212.3 (17.3)	212.5 (17.3)
Undisclosed	55	4.66 (0.37)	RM-trimmed data	5.26 (0.49)	4.45 (0.21)
Fatty Acid	48	39.29 40.55	Mode KDE Mode KDE	39.66 (0.12)	39.56 (0.1?)

Values in parenthesis are standard error of the mean (sem). $\text{sem} = \delta / \sqrt{n}$

¹⁰ Lothian and Thompson (2002)

¹¹ RM=Robust Mean, KDE= Kernel Density Estimate

The independent quantitative evaluation of these data were undertaken by the authors Lothian and Thompson [2002] using a series of statistical methods including bootstrapping and robust statistics, with and without truncation of extreme data or the second of alternative modes. The main disadvantage of utilising these methods where there are multiple modes is that the division or truncation of the data is subjective and, sometimes, arbitrary.

The main advantage of the Cofino BWE model is that:

- i) provides detailed quantitative statistics,
- ii) both the quantitative data and the graphical outputs are obtained in a single evaluation and
- iii) all data were evaluated without elimination of outliers, or by subjective truncation or separation of modes.

The extent of the level of agreement between the published data using the KDE, bootstrapping and robust statistics and with the Cofino BWE model described in this paper on a wide range of quite difficult datasets provides good corroborative evidence for both methods.

5. THE QUASIMEME LABORATORY PERFORMANCE STUDIES 1996-2001

The application of the Cofino model to the ongoing evaluation of the LP studies is dependent on being able to demonstrate its fitness for purpose. This has been achieved through the retrospective assessment of the wide range of types of data available from these studies, which has enabled the model to be fully examined and compared both with robust and conventional statistical analysis.

The QUASIMEME LP studies have been conducted for the measurement of chemical determinands in marine matrices as part of the external quality assurance of information to underpin national and international marine monitoring programmes, which include the Oslo and Paris Commission (OSPAR) and the Helsinki Commission (HELCOM).

Data from these studies were assessed at the outset of the project in 1992 using robust statistics [AMC 1989 a & b, Cofino & Wells 1994, Wells & Cofino 1997, de Boer *et al* 2000]. Robust statistics does not fully overcome the problems of skewed and tailing data, and cannot detect modality. The Cofino model was used as an alternative and overcomes the limitations of the robust statistics.

The determinands in this review cover nutrients in seawater and estuarine waters, trace metals in fish, shellfish and sediment, organochlorine pesticides, chlorobiphenyls and polycyclic aromatic hydrocarbons also in fish, shellfish and sediment. A list of each chemical or element is summarised in Table 4. Most of the determinands listed are mandatory for the marine monitoring programmes.

TABLE 4

The data was collected over 6 years, 1996 – 2002, from 12 exercises of the QUASIMEME LP studies, with the exception of MS3, where only 11 exercises were included in this assessment, and BT4, where data from only 5 exercises were available. There were 2–3 test materials in each exercise. A total of 2055 records were included. The number of participants reporting data for any determinand in a single test material ranged from 10–66.

Determinand Group	Determinands	Matrix	Group ID	Number of records	Maximum number of data for any study
Nutrients	Nitrite, Ammonia, Phosphate, Silicate	Seawater	AQ1	120	66
		Estuarine and Low Salinity Open Water	AQ2	160	62
Nutrients	Total nitrogen, Total phosphorus	Seawater	AQ1	48	43
		Estuarine and Low Salinity Open Water	AQ2	64	39
Trace Metals	Arsenic, Cadmium, Chromium, Copper, Lead, Mercury, Nickel, Zinc	Biota	BT1	200	44
		Sediment	MS1	200	48
Trace Metal Cofactors	Aluminium, Iron, Lithium, Manganese, Total organic carbon	Sediment	MS1	113	48
Chlorobiphenyls	CB28, CB52, CB101, CB105, CB118, CB138, CB153, CB156, CB180	Biota	BT2	216	41
		Sediment	MS2	216	43
Organochlorine Pesticides	pp' DDD, pp' DDE, op' DDT, pp' DDT, Dieldrin, HCB, α HCH, γ HCH, Transnonachlor	Biota	BT2	216	38
		Sediment	MS2	216	28
Polycyclic Aromatic Hydrocarbons	Benzo[a]anthracene, Benzo[b]fluoranthene, Benzo[e]pyrene, Benzo[g,h,i]perylene, Chrysene, Fluoranthene, Indeno[1,2,3 cd]pyrene, Phenanthrene, Pyrene	Biota	BT4	90	34
		Sediment	MS3	198	30

The retrospective evaluation was required to demonstrate

- i) The underlying normality in the data or a sub-set of the data.
- ii) The suitability of the BWE model to deal with bi or multi-modality.

- iii) That the BWE model provides the best estimate of the concentration of the determinand with the types of skewed, non-normal data that occur in such interlaboratory studies without trimming, elimination of outliers or any subjective pre treatment.

5.1 The Kolmogorov-Smirnov Test for Normality

The Kolmogorov-Smirnov (K-S) test for normality was used on each dataset to establish how many were normally distributed.

The K-S test, with the Lilliefors modification, has been applied to all data in this evaluation to determine if the null hypothesis of composite normality is a reasonable assumption regarding the population distribution of each linear dataset. The Lilliefors test is based on simulation, therefore the significance level is restricted to $0.01 \leq \alpha \leq 0.20$, which covers the critical values tabularized by Lilliefors. The Lilliefors test is a 2-sided test of composite normality with sample mean and sample variance used as estimates of the population mean and variance, respectively. The test statistic is based on the *normalised* samples whereby the Z-scores are computed by subtracting the sample mean and normalising by the sample standard deviation.

$$Z = (x_i - \bar{X}) / \delta$$

In this instance δ , the s.d., is derived from the dataset rather than an external standard, as in the QUASIMEME LP studies.

Of the 2055 datasets, only 20% passed the K-S test when all data were used (Table 5). The distribution of the percentage of data attributed to PMF₁, the first mode, and the proportion of K-S pass/fail is given in Figure 14a. Also included in this plot is the distribution of the percentage of data in the first mode of the NDA.

For the purposes of this paper the *heart-cut* of data taken has been defined as a range of data with the loci at the mean_{NDA}. The following ranges have been used:

Heart-cut-45 = mean_{NDA} \pm 45% of data

Heart-cut-35 = mean_{NDA} \pm 35% of data

and

Heart-cut-25 = mean_{NDA} \pm 25% of data

If the mean_{NDA} occurs at the 50 percentile then the *heart-cut-25* would be equivalent to the IQR. The criterion for selecting the *heart-cut* is given in the discussion of the BWE, above.

TABLE 5

Summary of the Test of Normality on the QUASIMEME LP studies (1996-2001) using the Kolmogorov-Smirnov test with the Lilliefors modification

	All data	<i>Heart-cut</i> - 45	<i>Heart-cut</i> - 35	<i>Heart-cut</i> - 25
K-S fail	1645	497	281	201
K-S pass	410	1494	1686	1691
Total	2055	1991	1967	1892
% pass	20	75	86	89
Percentage count of normality				
Normal for all selected data	Normal for 3 data ranges	Normal for 2 data ranges	Normal for 1 data range	Not Normal for all ranges
18	52	16	8	6

The Cofino Model: A Handbook to Evaluate Laboratory Performance Studies

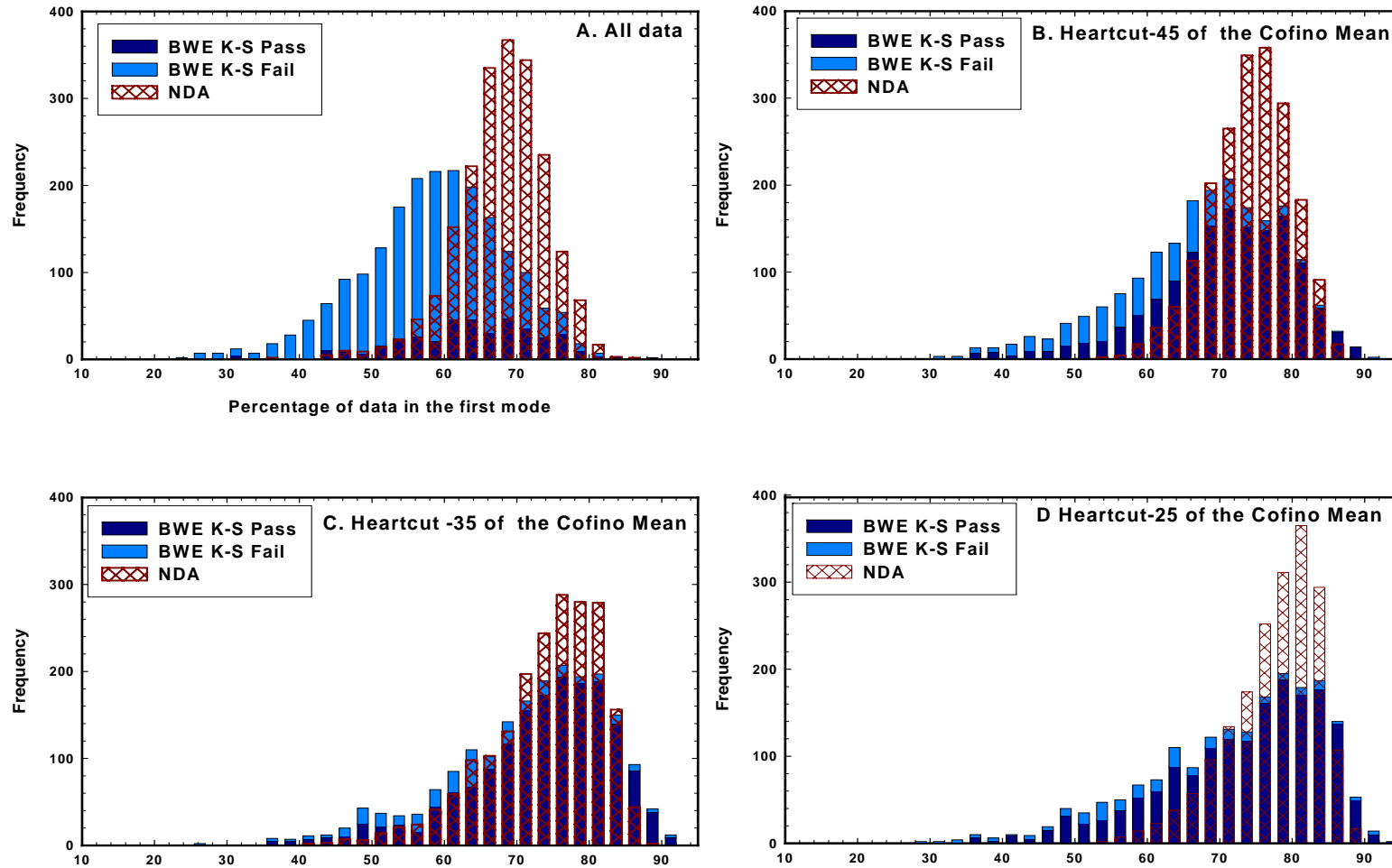


Figure 14. Distribution of the percentage of data attributed to the first mode using the Cofino BWE and NDA model and the ratio of the K-S pass/fail for the 2055 data sets of the QUASIMEME LP studies 1996-2001. The BWE histogram is stacked according to the ratio of data that passes or fails the K-S test for normality. A = All data, B = *Heart-cut* -45 i.e. Model mean \pm 45% of data, C = *Heart-cut*-35, D = *Heart-cut*-25.

The low percentage of K-S passes is primarily due to a combination of extreme values ($|Z| > 6$) and unsatisfactory values ($|Z| > 3$) which, for many datasets, give a positively skewed distribution. The mode of the % data_{BWE} is 62% and for the percentage data_{NDA} it is 70%.

By selecting the *Heart-cut-45* the K-S passes increase from 20% to 75%. The mode of the % data_{BWE} increases to 75% and to 77% for the NDA model (Fig. 14b).

A further reduction in the proportion of data in the BWE to *Heart-cut-35* and *Heart-cut-25* increases the K-S pass to 86% and 89% respectively and the mode of the % data_{BWE} and the % data_{NDA} to 77% and 80% respectively (Figure 14c & d).

The increase in the percentage of K-S passes from 20% to 89% for all data and the *Heart-cut-25* respectively confirms the assumption made that the data about the Cofino mean in most datasets is normally distributed.

As the *heart-cut* of the data is refined the mode of the percentage of data associated with the PMF₁ increases both for the BWE and for the NDA models, but more so for the BWE. The distribution of the % data_{BWE} becomes more skewed as the *heart-cut* of data taken from each dataset decreases. This occurs mainly from the more difficult datasets with bimodal or multi-modal distributions and a higher percentage of LCVs.

5.2 Comparison of Models using the QUASIMEME LP Data

The underlying normality of the QUASIMEME LP data can be visualised by comparing the mean and s.d. from the Cofino BWE model with those of the other models viz Cofino NDA model, robust statistics and normal statistics.

This comparison has been made by taking the ratio of the Cofino BWE model for both the mean and s.d., in turn, with the respective mean and s.d. of the other models. These ratios have been obtained for all data, and for the datasets taking the *Heart-cut-45*, *Heart-cut-35* and *Heart-cut-25* of each dataset.

The ratios for the 2055 datasets are summarised as frequency distribution curves (Fig. 15).

5.2.1 Means

The means for the Cofino BWE and the NDA model are very similar for all data and for each of the three *heart-cuts* (Fig. 15a), with a ratio of 1 for most data. Small differences occur with few data (NObs < 15) or more difficult datasets, but generally there is good agreement.

There is also a good comparison between the BWE and robust means when the *Heart-cut-25* of each dataset is taken (Fig. 15b). As more data are added back into the dataset with the *Heart-cut-35* and *Heart-cut-45* the differences between the BWE and the robust means become more apparent. For most cases the ratio increases as a result of the greater value of the robust mean. This clearly shows why it was previously necessary to trim the data (i.e. take a *heart-cut*) prior to estimating the robust means.

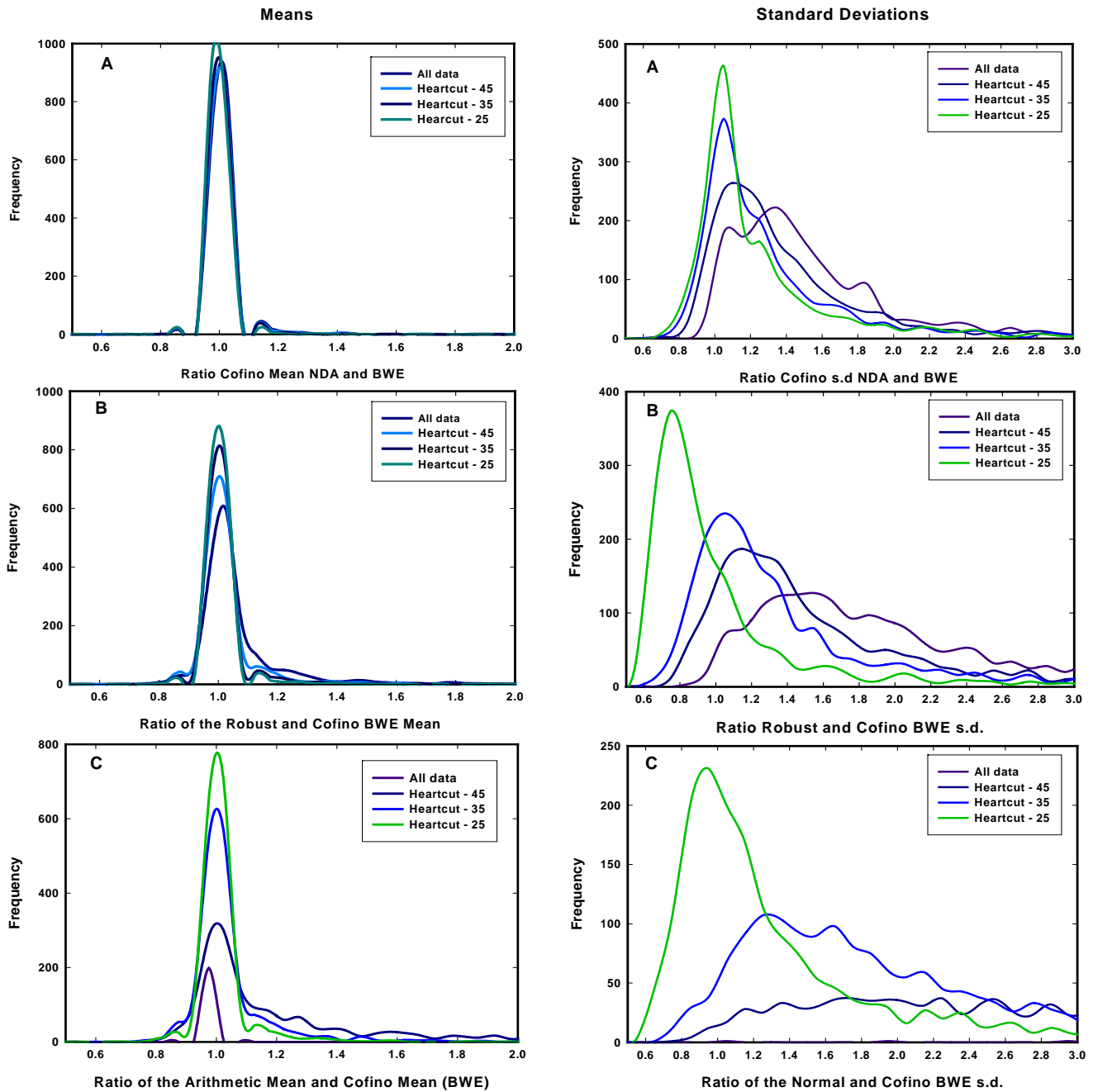


Figure 15. The distribution of the ratio of means and standard deviations for the datasets from the QUASIMEME LP studies (1996-2001) (n=2055)

A) Ratio of means $\text{NDA} / \text{means}_{\text{BWE}}$ and $\text{s.d}_{\text{NDA}} / \text{s.d}_{\text{BWE}}$

B) Ratio of means $\text{Robust} / \text{means}_{\text{BWE}}$ and $\text{s.d}_{\text{Robust}} / \text{s.d}_{\text{BWE}}$

C) Ratio of means $\text{Normal} / \text{means}_{\text{BWE}}$ and $\text{s.d}_{\text{Normal}} / \text{s.d}_{\text{BWE}}$

The Cofino NDA (A), robust statistics (B) and the conventional (normal) statistics (C) are compared with the Cofino BWE for all data and selected *heart-cut* of data.

There is a similar distribution for the comparison of the BWE and the arithmetic mean, except that the differences with the *Heart-cut-45* and all data are considerably greater (Fig. 15c). Differences in the means obtained by these two methods are to be expected with non-normal, skewed data. However, the importance of this relationship is shown in the marked increase in the agreement between the BWE and arithmetic mean as the *heart-cut* of data taken decreases to the *Heart-cut-35* and then to the *Heart-cut-25*. This, again, underlines that within most of the datasets there is a *heart-cut* which behaves as if it were normally distributed.

5.2.2 Standard deviation (s.d.)

A similar pattern occurs for the s.d. with each model except that in most cases the magnitude of the s.d. for the robust and the normal statistics is higher than for the BWE. Although a considerable number of datasets have a similar s.d. for the BWE and the conventional statistics when the *Heart-cut-25* of data is taken, there is also clearly a difference for other datasets. However, this is to be expected since the s.d. for the two models have a different basis. The s.d. for the BWE and the conventional statistics are only the same IF the datasets are the same and the data are normally distributed.

The main contribution to the s.d. _{BWE} comes from data that overlap the most (have a high level of agreement). These data also have the highest coefficients in the model's calculation. Although there is a contribution from all data to the s.d._{BWE}, those values that are very different from the expectation value of PMF₁ have very little effect on either the mean and the s.d._{BWE}. Therefore it is possible to make an empirical comparison by selecting the data which contribute the most to the mode of PMF₁.

Two examples are given in Table 6. Benzo[b]fluoranthene, which is discussed later with respect to bimodality (Fig. 16A), has 41% of the data that comprise the mode of PMF₁. This percentage is equivalent to the 9 values in good agreement that straddle the mean. A comparison of the BWE and the arithmetic means and s.d. for these 9 values show relatively good agreement (Table 6).

TABLE 6

Comparison of the mean and s.d. of all data and selected data around the Cofino mean using the BWE model and normal statistics

	BWE		Normal	
	mean	s.d.	mean	s.d.
Benzo [b] fluoranthene				
All data n=21	628.5	29.3	700.3	472
Selected data n=9	623.9	24.5	624.1	17.1
Cadmium				
All data n=37	2.62	0.56	7.89	10.5
Selected data n=11	2.89	0.66	2.89	0.44

A second example is given for cadmium in flounder muscle (Fig. 17). The Cofino BWE model mean is 2.62 ± 0.56 (n=37) while the arithmetic mean is 7.89 ± 10.5 . By selecting the data that are in good agreement there is much close agreement between the BWE and arithmetic mean.

Eliminating outliers and using normal statistics or calculating the robust descriptors with or without removing extreme values have all been methods by which the best estimate of the true mean and the s.d. have been attempted. In using the Cofino BWE model it is possible to achieve this without the need to identify and eliminate outliers, or to down-weight, trim or select data. The model provides a mean and s.d. that is fully traceable and defines the contribution that each data point makes to these statistics.

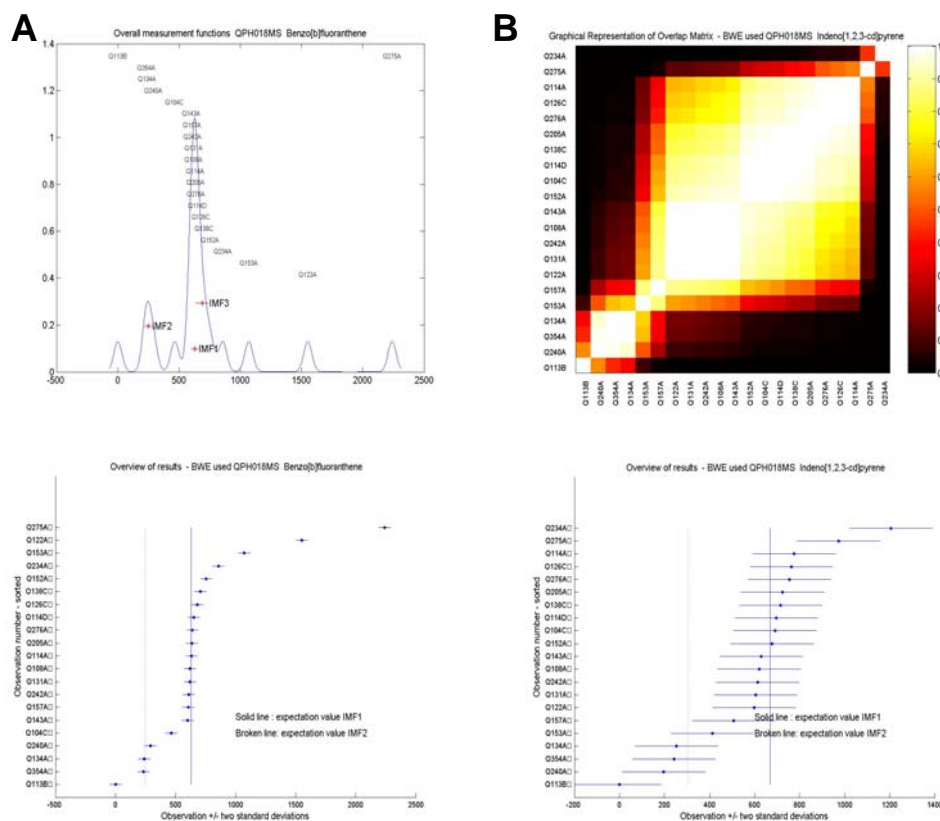


Figure 16. Overview results, measurement functions and Kilt plots for PAHs showing multi-modality (A) Benzo [b] fluoranthene (B) Indeno [1,2,3-cd] pyrene

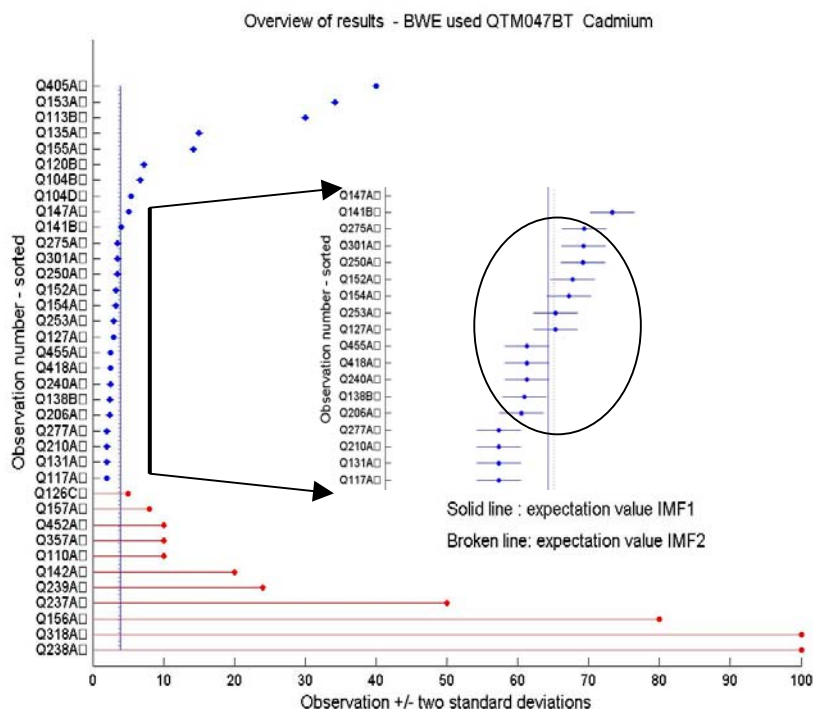


Figure 17. Overview of results for cadmium in flounder showing the data in relatively good agreement (circled). These data have the highest contribution to the values associated with the main mode (PMF₁). The red lines and circles represent the LCVs.

5.3 Number of Observations and Percentage of Data in the First Mode

The QUASIMEME LP studies stipulate a minimum of 10 participants before initiating any study. For most studies the numbers are well in excess of 10, however not all determinands are measured and there are sometimes nil returns which does mean that NObs in some datasets are < 10. Of the 2055 datasets only 28 (1.4%) have NObs < 10. The distribution of the NObs for each group of determinands in each study is given in Figure 18. The distribution for the nutrients is bimodal because the number of laboratories returning data for total nitrogen and total phosphorous are always less than for the other nutrients. Likewise for each of the other groups there are some determinands which are only analysed by a few laboratories. However, it is clear from the evaluation of the data that the non-normality does not come specifically from the data with a low number of NObs.

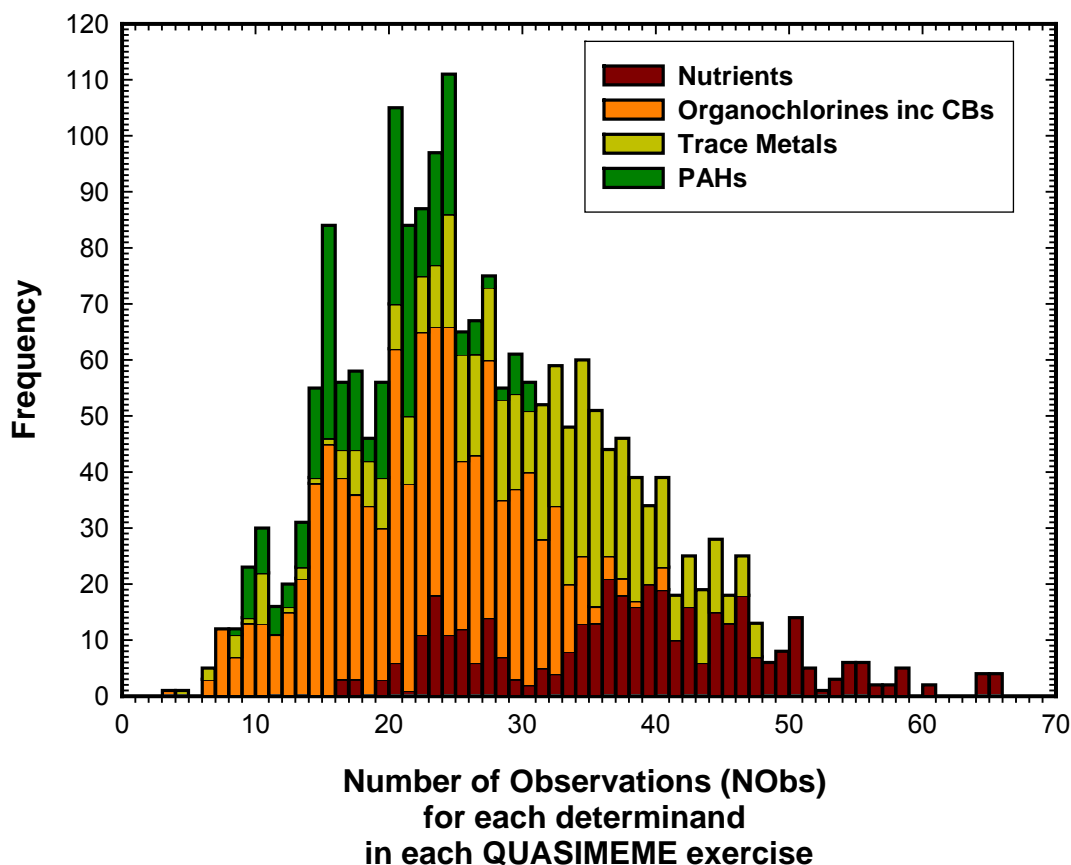


Figure 18. Distribution of the NObs for each determinand, by determinand group QUASIMEME LP studies 1996-2001.

The relationship between the % data B_{WE} associated with the PMF_1 and the NObs in that dataset are given in Figure 19. The data are plotted in two groups: those that passed the K-S normality test and those that failed.

In general, those datasets that passed the normality test tend to have a higher percentage of data associated with PMF_1 . This is to be expected. Also as the NObs increase there is a tendency to have more data associated with the first mode, PMF_1 . However, it is not a simple relationship, as the scatter diagram (Fig. 19) suggests. There are some data with over 30 observations which can have as little as 30–35% of the data associated with the first mode. This would seem unusually low until details of the data are reviewed.

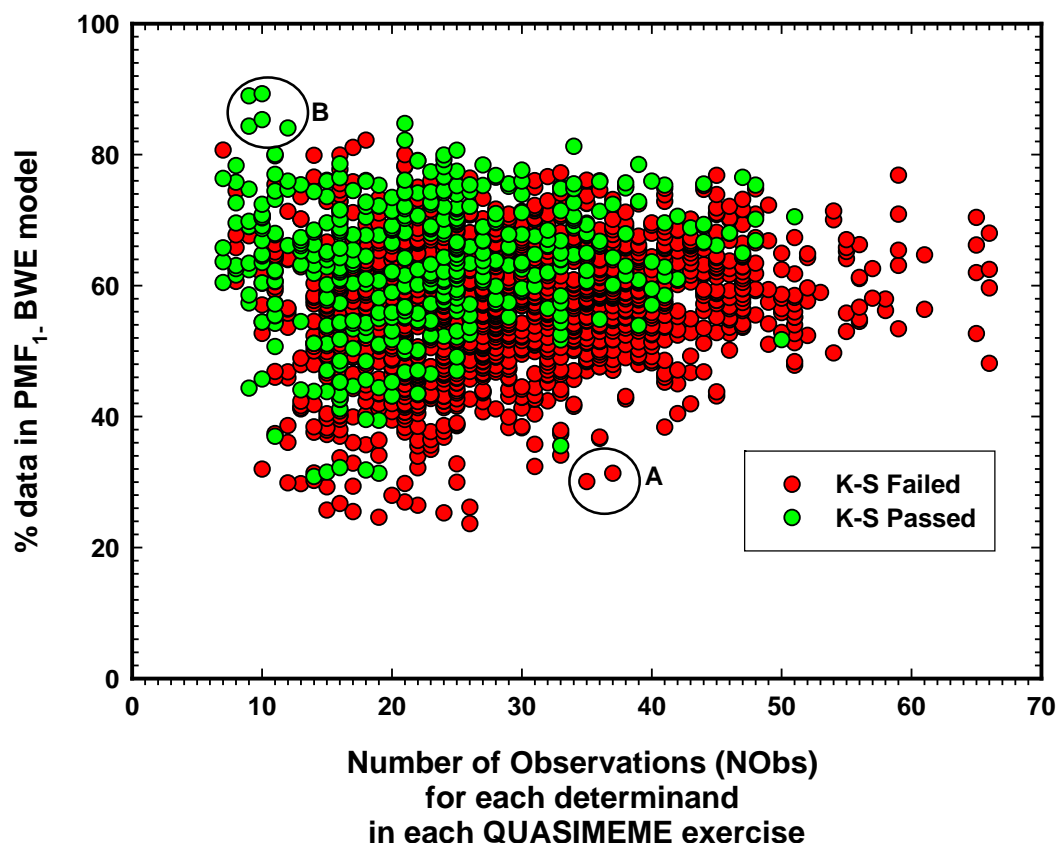


Figure 19. Relationship between the %data attributed to the first mode (PMF₁) and NObs for each determinand in QUASIMEME LP studies 1996-2001

The two datasets marked “A” in Figure 19 are both for cadmium in fish muscle. Cadmium in fish muscle is normally low, at ca 1 – 10 µg/kg. The overview of the results for one sample analysed for cadmium are ranked and plotted in Figure 17. Of the 37 sets of data, 11 are LCVs with a maximum of <100 µg/kg. The next 16 values range from 2–3.5 µg/kg with a further 8 increasing in value to 40 µg/kg. In the heart of the data there are 12 values, around 30–35% of the data, which are in relatively good agreement and form the basis of the data that comprise the main mode, PMF₁. Many of the LCVs are too high to be included in the calculations using a rectangular pdf as only 2 LCVs are less than the Cofino mean of all numerical values. The summary statistics for the evaluation of cadmium in fish muscle are given in Table 7.

Neither the robust statistics nor the normal statistics are able to cope with these distributions.

In contrast, the four values (B) in Figure 19 from the measurement of PAHs in mussels have 12, 10 x 2 and 9 observations in the dataset. Over 80% of these data are included in the mode of PMF₁ and each dataset is normally distributed.

TABLE 7

Cadmium in mussels

Method	n	Mean	s.d.	% of data in mean
Cofino BWE – no left censored	26	2.62	0.56	33.4
Cofino NDA – no left censored	26	3.10	1.65	67.7
Rectangular pdf for left censored				
Cofino BWE – 11 left censored	37	2.59	0.49	32.3
Cofino NDA – 11 left censored	37	3.05	1.59	67.6
Robust – no left censored	26	4.44	2.85	-
Normal – no left censored	26	7.89	10.52	-

Therefore datasets which form part of the tail in the distribution of the % data _{BWE} (Figure 14a) are likely to be data which are heavily skewed by positive extreme values and / or by a large number of LCVs. This frequently occurs with determinands whose concentration is close to the LOQ. In such cases false positive values can occur more easily through contamination or using relatively contaminated reagents. A greater number of LODs are also reported at these concentrations. The LCVs included in the Cofino model are discussed later.

5.4 Bimodality in QUASIMEME Datasets

In addition to the extreme values in any dataset, skewed distributions are common place due to a wide range of factors such as contamination or calibration. These can be readily handled with using either of the Cofino models. The other main pattern to resolve both quantitatively and qualitatively is the bimodal or multimodal distribution.

Many factors contribute to bimodality and it is therefore important to separate and identify the data associated with each mode and, where possible, the source of the differences. The sensitivity of the BWE model to both the degree of separation and the balance of data in the two modes has already been discussed.

However, for routine LP studies it is important to have a simple, automatic screening process to detect bimodality in data prior to subsequent investigation. It is also important that the method of evaluation can make the distinction between genuine modality with a root cause, and artefacts.

Unless there is a significant separation of values most of the data (ca >90%) are accounted for by the first two modes, PMF₁ and PMF₂. This means that for significant bimodality to occur the data associated with PMF₂ would need to be distinct from the data associated with PMF₁. Since the data associated with both PMF₁ and PMF₂ have known mean and s.d. from the Cofino model, it is possible to compute a simple Student's t test as an indication of bimodality. As an approximate guide, a Student's t test value >2 was used to denote a significant difference between the data associated with PMF₁ and PMF₂. These datasets required further investigation.

Of the 2055 datasets, 1743 had Student's t values <2, leaving 260 with potential bi- or multi modality. The ranked Student's t values along with the % data associated with PMF₁ / PMF₂ are given in Figure 20. As the Student's t value increases in magnitude the ratio of the data associated with PMF₁ / PMF₂ declines, indicating an increasing proportion of data occurring in the second mode and the significance of the bimodality.

The 260 datasets with a Student's t value >2 were ranked in order of magnitude and collated by determinand. Of the 24 test materials for each determinand group distributed during 1996 – 2001, there were 16 (60%) occurrences of aluminium having a bimodal distribution. This specific case is discussed in more detail later.

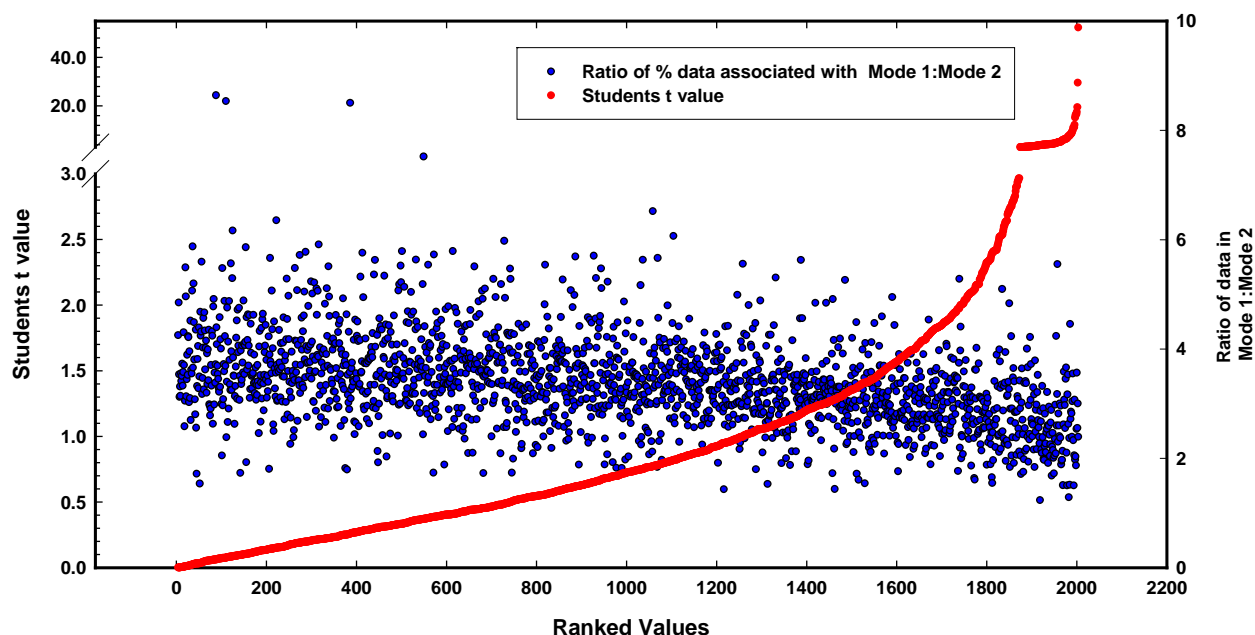


Figure 20. The Student's t test value of each data set ($n=2055$) in the QUASIMEME LP studies (1996-2001) ranked in order of magnitude and the ratio of the % of data associated with mode 1 and mode 2 of the data. As the data inclines toward bimodality the value of the Student's t test value on mode 1 and mode 2 will increase and the distribution of the data in mode 2 will increase. A Student's t test value of > 2 would indicate bimodality.

The next most numerous group of bimodal data were the PAHs: benzo[a]anthracene, 8 (33%), benzo[b]fluoranthene, 10 (42%), benzo[g,h,i]perylene, 11 (40%) and indeno[1,2,3-cd]pyrene, 11 (46%). These cases were spread equally between both sediment and mussel test materials. Having such a frequent occurrence of bimodality for these PAHs, there was a high probability of systematic differences in the data based on methodology / technique used.

The multi-modality is clearly evident in the two examples of benzo[b]fluoranthene and indeno[1,2,3-cd]pyrene in marine sediment (Figure 16 B). For benzo[b]fluoranthene there are a group of 10 – 11 laboratories in relatively good agreement, 5 – 6 positively skewed values with no overlap, and a second group of 3 at a lower concentration (Fig. 16 A). The low outlier Q113B reported a value of 0.72. The units of the determinand should have been in $\mu\text{g/kg}$ rather than mg/kg . This is a clear example of good analysis and mistaken reporting.

For indeno[1,2,3-cd]pyrene, there is a similar bimodal distribution, but with fewer positively skewed values. The bimodality is clear both on the overview of data and on the Kilt plot. Laboratory Q113B appears to have made the same error for this compound. The cluster of laboratories in the second mode is similar for both compounds.

Looking at the rescaled sum of z-scores (RSZ) [Thompson & Wood 1993] it was possible to detect that three of these laboratories had a long term negative bias for these compounds, independent of the type of marine sediment (Table 7).

The Student's t value has effectively been used as an initial screen for bi- or multi- modality that can focus further investigation. The Cofino BWE model provides sufficient structured detail to identify the most salient features and gives sufficient evidence of the actual heterogeneity that can exist in a dataset.

TABLE 7

Outliers in the measurement of PAHs in marine sediment

Benzo[b]fluoranthene	Indeno[1,2,3-cd]pyrene
Q113B	Q113B
Q354A	Q354A – S
Q134A – S	Q134A – S
Q240A – S	Q240A - S
Q104C	

S = long term negative bias identified by re-scaled sum of z-scores in studies 1996 – 2001.

5.4.1 Examples of bimodality

Three further examples of bimodality are taken from the QUASIMEME LP studies to demonstrate the structural information that becomes available when applying the Cofino BWE model. The ranked overview of results, the summed probability density functions and the Kilt plot is given for the three examples in Figure 21.

5.4.1.1 Chromium in Cod Liver

The level of chromium in biological tissue is generally low and contamination can easily be a source of error unless an ultra clean environment is used. The data are clearly bimodal with an expectation value of PMF₁ around 0.11 µg/kg (Fig. 21a). The digestive media included hydrochloric acid, aqua regia, nitric acid, nitric acid with peroxide and five others unspecified. There was no clear preference for either group, however, most of the laboratories used nitric acid and with or without peroxide (19). Six laboratories used an open heating system for digestion and five of these were within the upper 50% of the data with three being in the second mode. Microwave digestion and the pressure bomb formed most of the data in the first mode.

The method of detection also separates approximately into the two modes with the main mode, PMF₁, comprising of atomic absorption spectroscopy (AAS), electrothermal atomisation (ETM) with and without a chemical modifier to the Zeeman background correction and the second mode, PMF₂ which was primarily Inductively Coupled Plasma (ICP) with atomic emission spectroscopy (AES) or mass spectrometry (MS).

Clearly, there are potentially competing factors in the methodologies, which give rise to the bimodality and contamination from digest in open vessels cannot be ruled out.

5.4.1.2 Chlorobiphenyl (CB105) in Mussels

CB105 is amongst the more difficult congeners to completely separate chromatographically from other CBs [Wells, 1999] in particular CB132. The higher values in the second mode (Fig. 21b) are most likely to be associated with an insufficiently cleaned up sample, degrading or inappropriate chromatographic column selection and/or optimisation.

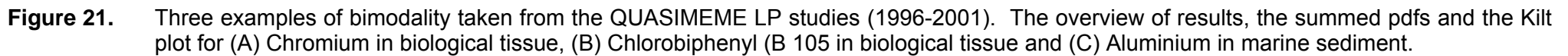
The data associated with the expectation value (PMF₁), has a mean of 0.27 µg/kg while the second mode is about 2.5 times higher at 0.64 µg/kg. In most cases where there is clear information on likely interferences, the best estimate for the assigned value is often the mode with the lower value. However, this does not always need to be the case and careful judgement with supporting chemistry is essential.

5.4.1.3 Aluminium in Sediment

The selection of digestion methods affects the recovery of refractory elements such as aluminium from clay mineral sediments. A range of strong acids are normally used, with hydrofluoric acid (HF) being amongst the most effective at breaking the aluminosilicates and releasing the aluminium. Other digestive techniques used are aqua regia (AR), nitric acid (HNO₃) as well as non-destructive techniques (NDT) such as X-ray fluorescence. Most laboratories currently use HF as the preferred digestive 'wet' technique.

The results for aluminium (Al) in the marine sediment (Fig. 21c) show three obvious modes in the data. The highest mode, around 7.5% Al is probably a result of two laboratories obtaining a particularly high value. The main mode, PMF₁, has a concentration of 4.95% and consists almost exclusively of digestion methods involving HF or non-destructive techniques. The lower mode, PMF₂ at 1.6% Al includes methods using aqua regia (2), nitric acid (2), nitric acid and peroxide, and HF (4). The non-HF, wet techniques clearly do not extract all of the aluminium, but the use of HF *per se* does not always guarantee complete recovery of the element. Again, in this instance the most homogeneous and reliable assigned value was based on the main mode, PMF₁.

45



6. LEFT CENSORED VALUES

Left Censored Values (LCVs), “less than” values are a common feature in LP data, especially when the concentration of the determinand is at low concentrations relative to the limit of determination. There have been a number of different approaches to incorporating these values into the data assessment and treating them as part of the whole evaluation [Gilbert 1987, Gilliom & Helsel 1986, Helsel & Gilliom 1986, Helsel & Cohn 1988, Kuttatharmmakul *et al.*, 2001]. This is a relatively important aspect of the quality assurance of measurements at these low levels since the LCVs may well, in many cases, be a truer reflection of the actual concentration rather than the positive values which may only be representative of the measurement of contaminated samples.

The Cofino model has been extended to include LCVs. A summary of the mathematical model and the extension to include the LCVs has been given earlier in this paper and published elsewhere [Cofino *et al.*, 2004 in press].

All numerical data and LCVs are included in the datasets used by the Cofino model. The only constraint that has been imposed has been to limit the magnitude of the LCVs to the mean \bar{NDA} . The higher LCVs, whilst encompassing the Cofino mean, often cover a large range and contain little information on the actual level of the determinand. In addition the rectangular pdfs of these LCVs would interact with the normal distribution pdf of the outlying observations, with a result that the Cofino mean would be falsely elevated. High LCVs are frequently associated with methods that have insufficient sensitivity for the measurement at the concentration of the determinand.

LCVs that are less than the Cofino mean are included in the model calculation. Where there are a sufficient number of these values it will influence the final consensus value by lowering it with respect to the mean of numerical values only. This effect is what should be expected and will often provide a more reliable estimate of the true value. This effect can best be explained by the example of pp' DDT (QOR055BT) in mussels (Fig. 22).

The pp' DDT occurs at low concentration, close to, or below the LOQ. Many data for this compound are reported as LCVs. The data that are numerical values can often be associated with contamination or insufficient sample clean-up procedures.

Therefore it is more realistic for the consensus value resulting from the Cofino BWE model to be influenced by LCVs less than the mean rather than high and possible false positive values.

When the LCVs are not included in the calculation there are only 13 values. The distribution of the data is bimodal with 38% of the data associated with the high mode centred around $1.15 \mu\text{g/kg}$. There are two extreme values at 4 and $8 \mu\text{g/kg}$ respectively. The lower mode is centred around $0.05 \mu\text{g/kg}$ and is similar in magnitude to 9 of the LCVs. By including the LCVs the main mode now occurs at the lower mode of $0.054 \mu\text{g/kg}$ which is consistent with both 9 LCVs and 5 of the numerical values.

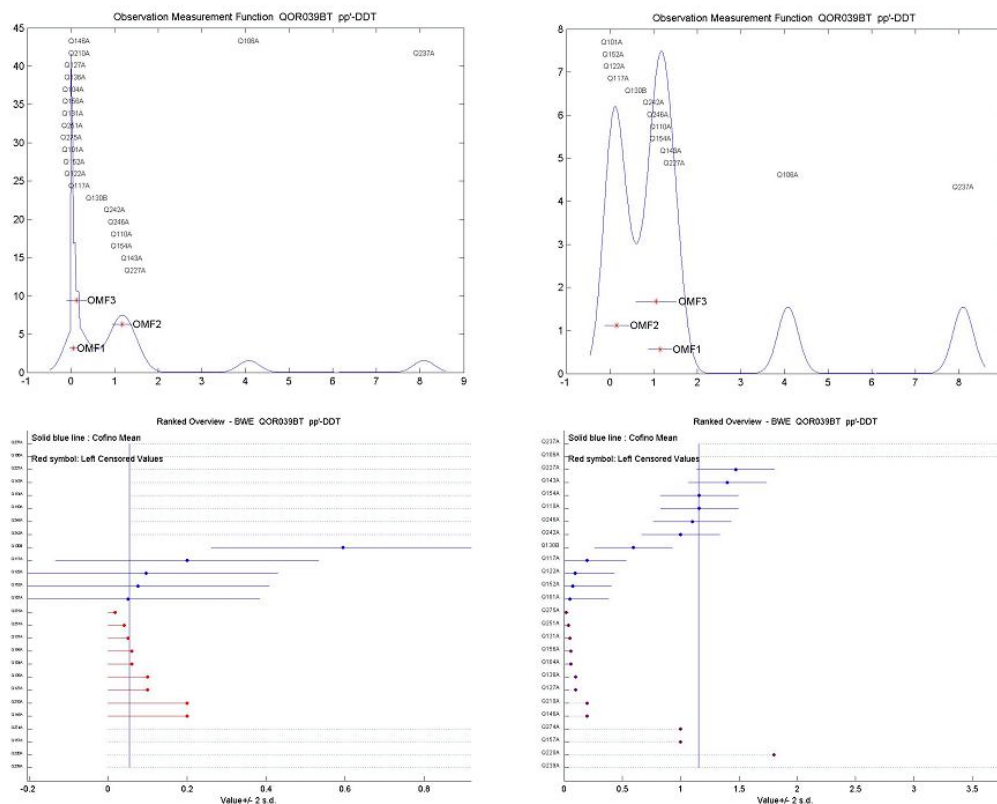


Figure 22. The overview of the results for pp' DDT with and without the restriction of limiting the LCVs to the Cofino mean of the numerical values.

Of the 2055 datasets within the QUASIMEME LP studies (1996-2001), 881 contained LCVs. The Cofino BWE model was used to calculate the summary statistics with and without the LCVs to examine what effect the inclusion of these data had on earlier assessments.

A plot of the % difference between the means calculated with and without LCVs, ranked in increasing absolute magnitude, is given in Figure 23. The number of LCVs as a percentage of the total NOBs for each dataset is also plotted.

Just over 43% of the 2055 QUASIMEME datasets have LCVs. Of these 881 datasets, 707 have means which differ by <5% by including the LCVs, 769 differ by <10% and 108 differ by >10% (Fig. 23). The distribution of LCVs by determinand groups is given in Table 8.

The OCPs in biota and sediment have the highest number of datasets with LCVs and they also have the highest percentage of LCVs in each dataset. In addition, the inclusion of these data has a significant affect on the expectation value calculated by the model. The two other groups that are affected are the CBs and the trace metals in biota, primarily from cadmium and lead that generally occur at very low levels in fish muscle tissue.

TABLE 8

Summary of the number of data sets containing LCVs and the number of data sets where the mean value differs by more than 10% with and without these LCVs.

Determinand Group	NObs data with LCVs	NObs data with LCVs Means _{LCV} / Means _{ALL} >5%
Nutrients in seawater	36	0
Total N and Total P in seawater	9	0
Chlorobiphenyls (CBs) in biota	133	9
Trace metals in biota	124	14
Organochlorine pesticides (OCPs) in biota	179	76
Polycyclic Aromatic Hydrocarbons (PAHs) in biota	33	2
Chlorobiphenyls (CBs) in sediment	32	5
Trace metals in sediment	134	1
Organochlorine pesticides (OCPs) in sediment	194	64
Polycyclic Aromatic Hydrocarbons (PAHs) in sediment	7	0

What clearly emerges, for some determinand groups, is the importance of including the LCVs in the whole assessment. It is also important for the assessment of laboratories that have reported LCVs. Clearly a Z score based on a specific value would be inappropriate for LCVs, but a flag for S = satisfactory would be appropriate where the value of the LCV was within the equivalent boundary of $Z < 3$. LCVs greater than the equivalent of $Z=3$ should be flagged as unsatisfactory.

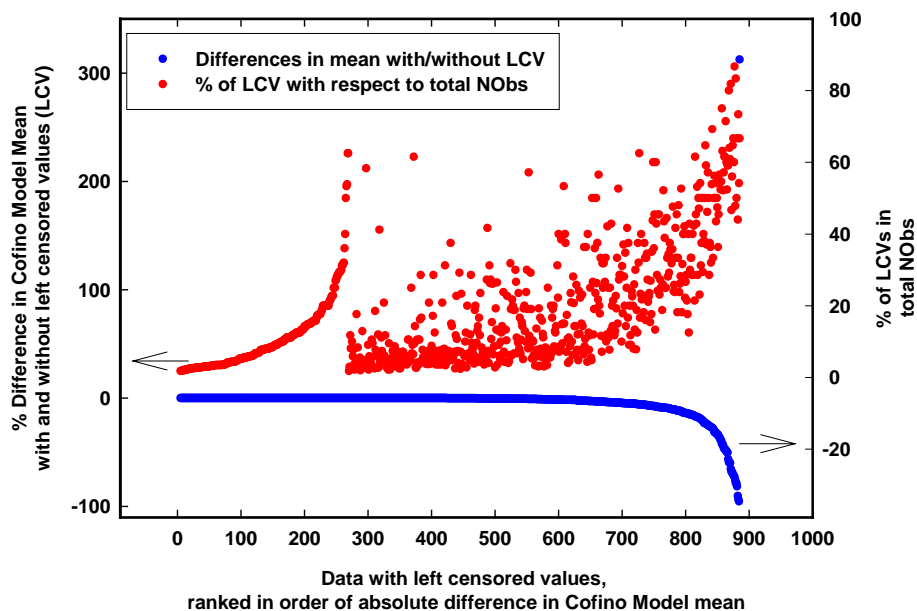


Figure 23. The percentage difference in the Cofino BWE mean with and without LCVs ranked in order of increasing absolute difference and the percentage of LCVs with respect to the total NObs.

7. SELECTING THE APPROPRIATE MODEL

The Cofino model has four different methods of operation that can be used calculate the population characteristics of the dataset. The selected model is required to be of sufficient sensitivity to distinguish the key features of the distribution without being unduly sensitive and providing false modes or spikes. Equally the model should be sufficiently responsive to provide important structural information.

The main methods available use:

- i) The *within-laboratory* variance from participants, either as replicate data or with a standard deviation provided.
- ii) The NDA model which bases the $s.d._{within}$ on the whole dataset.
- iii) The BWE model which takes a ± 2 s.d. *heart-cut* about the first mode and establishes the $s.d._{within}$ from the data in the *heart-cut*.
- iv) The KDE, which is a qualitative tool and bases the estimator, h_{opt} on the inter quartile range of all data.

Two examples have been selected to demonstrate the relative merit of each method. The first is taken from the UK Food Standards Agency (FSA) Proficiency Testing Scheme on Genetically Modified Organisms (GMO) and one from the International Atomic Energy Agency (IAEA) Interlaboratory Studies on Trace Metals in Sediment.

The output of the four models is given in the first example of GMO in soya in *Dutch Biscuit* (Fig. 24).

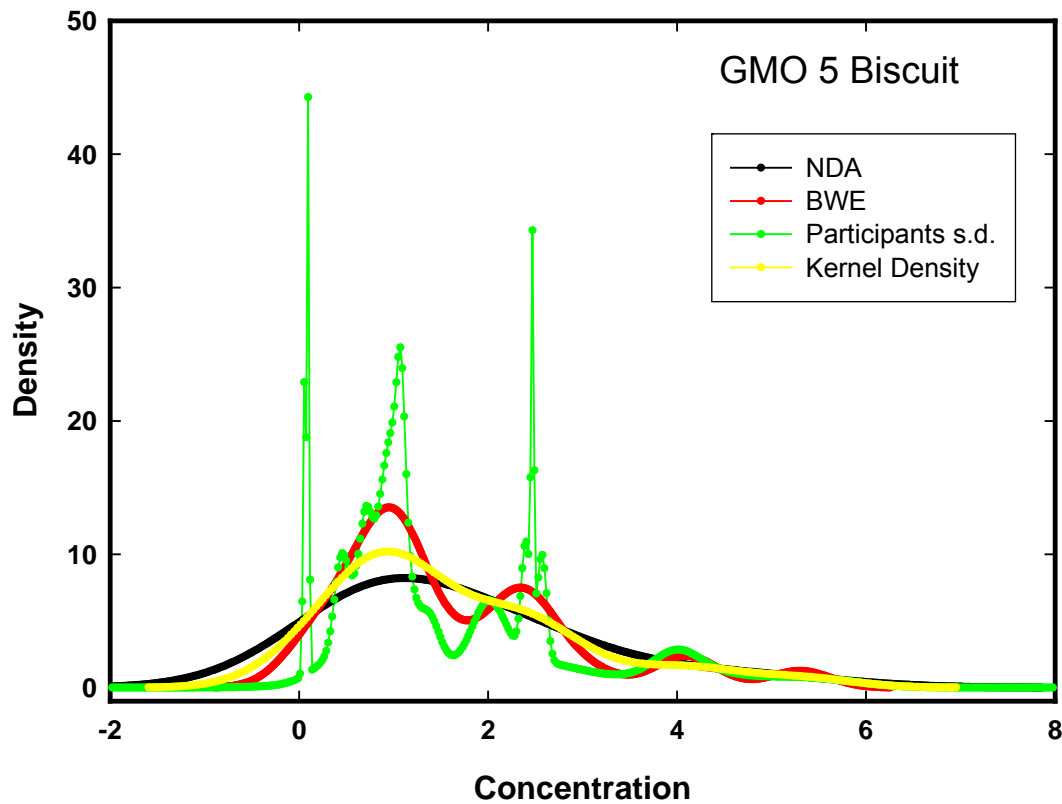


Figure 24. The overall measurement function for the data from the analysis of *Dutch Biscuit* for GMOs in the soya fraction of the test material.

Both the KDE and the Cofino NDA models provide an oversmoothing in the output with the main mode at *ca* 1% GMO and a small hump at 2.2%, tailing to 6%. When the participants' variance is used two main features become apparent. There are three spikes at 0.5%, 1% and 2.5% and smoother series of *modes* at 1%, 2.1%, and 4%. The spiking is caused by the data from individual laboratories or small groups of laboratories that have a relatively small s.d._{within}. The relatively high precision of these laboratories also means that the data has less overlap with neighbouring values. The lack of overlap between data results in a *mode* being created from these individual values, which has a very narrow bandwidth. This results in spiking superimposed on the overall distribution.

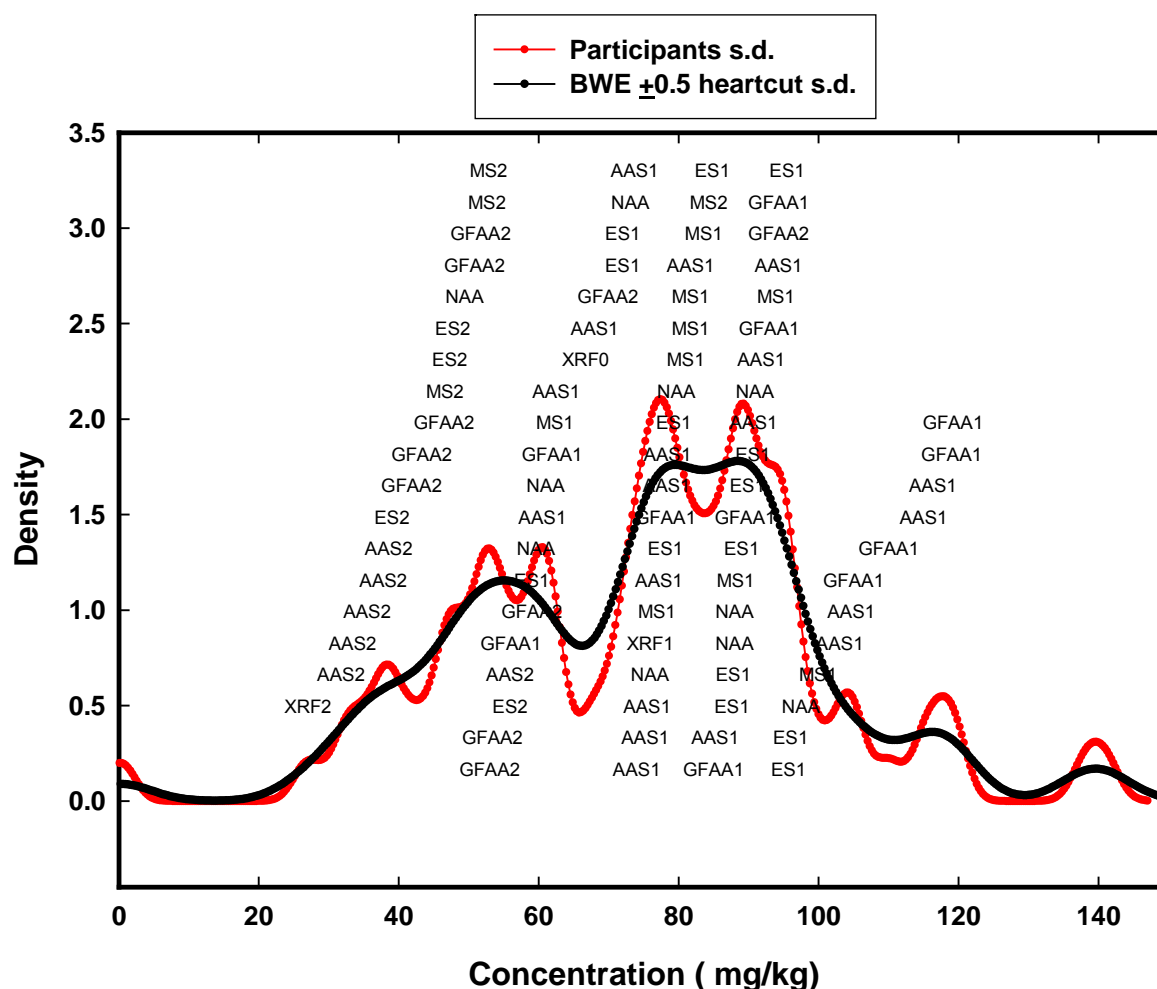


Figure 25. IAEA 405 Chromium in Sediment giving the overall measurement function comparing the profile obtained with the participants s.d._{within} and the BWE using ± 0.5 s.d. as the *heart-cut*. Superimposed on this are the method codes from the laboratories.

AAS1, AAS2,	Flame Atomic Absorption (AA) Spectrophotometry with / without HF
GFAA1, GFAA2	Graphite furnace AA with / without HF
MS1, MS2	Inductively coupled plasma mass spectrometry (ICPMS) without / with HF
ES1, ES2	Emmission spectrometry with / without HF
XRF1	X-ray fluroescence
XRF	X –ray fluroescence
NAA	Neutron Activation Analysis

The second example is drawn from the analysis of chromium in sediment (IAEA 405)¹². The s.d._{within} for each laboratory was also provided from replicate measurements of the sediment. Like most interlaboratory studies that request repeated measurements, the range of the s.d._{within} was, for most data, closer to the repeatability rather than the long term reproducibility (mean r.s.d._{within} = 2.25%, range 0.1%-13.25%).

The PMF obtained using the Cofino model and the participants s.d._{within} provided a considerable amount of structure to the profile. This profile was mirrored when the BWE was also used with a *heart-cut* of ± 0.5 s.d. The optimum *heart-cut* used is normally ± 2 s.d.

The main consideration in the interpretation of this profile was to determine whether the modes present were artifacts or whether they represented real analytical differences. The analytical methods used were coded with the data and an overall summary made based on digestion method and the final detection instrumentation.

It is well known that chromium, being a refractory element, requires full digest of the fluorosilicate to liberate all of the interstitial elements. The main digestion method for this is to use hydrofluoric acid (HF) as part of the acid mixture. The acids used in this interlaboratory study were characterized into four groups viz with and without HF and without and with oxidizing agent, perchloric acid or hydrogen peroxide.

There was little distinction between the groups with and without the oxidising agent, but, as expected, there were significant differences in the data generated using a digest without (1) and with (2) HF (Fig. 25). The data were separated on this basis and further divided according to the final detection instrumentation. The instrumental groups were Atomic Absorption (Flame) Spectrophotometry (AAS), Graphite Furnace Atomic Absorption Spectrometry (GFAAS), Inductively Coupled Plasma Mass Spectrometry (ICPMS), Emission Spectrophotometry (ES), X-Ray Fluorescence (XRF) and Neutron Activation Analysis (NAA).

Data were labeled according to the group of instruments used and by the digestion method, without / with HF. Data generated from digests without HF ranges from 22-60 mg/kg, with a few higher values. The data generated with HF digests ranges from 60-120 mg/kg. Within these groups the flame AAS systematically gave lower results than GFAA, ICPMS ES or NAA. These differences in methodology are summarized in Figure 25 and clearly demonstrate why these values are spread over such a wide range. However, as the data are method coded, the Cofino BWE model was able to demonstrate the method-related structure of the data and explain the much of the modality in the distribution of the overall measurement function.

8. DATA ASSESSMENT AND Z SCORE PLOTS

The Z score plots are a ranked representation of the data as a bargraph, normalised to the assigned value and the target s.d., such that

$$z \text{ score } (Z) = (\chi_i - \mathbf{x}) / \sigma$$

Where χ_i is the individual value, \mathbf{x} is the assigned value and σ is the target s.d. In some schemes the target s.d. is derived from the dataset, while in others it is set as an external standard of achievement. In the former, 95% of the data would achieve values of $|Z| < 2$.

¹² Data for this example was kindly provided by Dr Eric Wyse and Dr Stephen de Mora, IAEA, Marine Environment Laboratory, Monarcho.

QUASIMEME applied external standards for the performance requirements set by the Scientific Assessment Group.

The ISO 43 [ISO 1997] for laboratory performance study assessments provides three levels of evaluation. A z score of $|Z| < 2$ is classified satisfactory, $2 > |Z| < 3$ is questionable and $|Z| > 3$ is unsatisfactory.

QUASIMEME has added a further category of $|Z| > 6$ for assessment purposes. These are extreme values and may be regarded as not belonging to the population. They may be caused by the use of incorrect units, dilution / concentration or calculation errors rather than intrinsic errors in analytical methodology. Factors of 2,5,10 or 1000 are common for this type of extreme values.

In addition to the regular assessment for real values, QUASIMEME has introduced a further two categories for LCVs.

Until now LCVs have only been recorded and not evaluated. However, now that the model includes these values in the overall data distribution and calculation of the Cofino mean it is important to provide a more formal assessment of these values. This assessment is especially needed where the number of LCVs constitutes a significant fraction of the dataset.

Although LCVs $>$ mean of numerical values are excluded from the estimation of the assigned value all LCVs are assessed. The basis of the assessment is to confirm (or otherwise) that the range covered by the LCV is consistent with the assigned value. A LCV (e.g. < 0.1) was considerably smaller than the assigned value (e.g. 1) would imply that the method should have detected the determinand. This is not consistent with the assigned value.

Likewise a LCV which was much greater than the assigned value would imply that the method used was not sufficiently sensitive to detect the determinand, compared with other results.

The upper and lower boundaries selected are related to the target performance set by QUASIMEME to provide a common level of assessment. These boundaries are equivalent to $|Z| < 3$, where Z is the equivalent magnitude to the Z score and derived using the same formulae, as given above.

The actual value included in the calculation is 50% of the LCV. The rationale for this is simply taken from the distribution characteristics of the LCV, which is regarded as a rectangular probability distribution.

When this compared with the normal distribution characteristics of a real value, it is clear that the actual reported value is at the 50 percentile of the probability distribution (Figure 26).

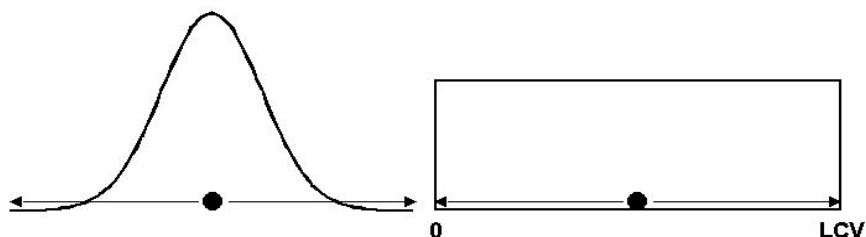


Figure 26. The normal probability distribution of a numerical values and the rectangular probability distribution of LCVs. The actual LCV reported in the upper boundary i.e. “less than” of the distribution that ranges from 0 to the LCV.

The two categories that have been assigned to the LCVs are:

- [1] **Consistent with the assigned value.** LCVs which fall into the category *consistent* have a value $LCV/2$ which is $< |Z| \leq 3$.
- [2] **Inconsistent with the assigned value.** LCVs which fall into the category *inconsistent* have a value $LCV/2$ which is $> |Z| \leq 3$.

The categories **consistent** and **inconsistent** are separate from **satisfactory** and **unsatisfactory** which are reserved for the assessment of numerical data.

9. ACKNOWLEDGMENTS

The authors would like to acknowledge the Food Standards Agency for their financial support during the development of the applications of the Cofino model, particularly in relation to the GMO proficiency testing data, and the Royal Society of Chemistry Analytical Methods Committee Statistics Group for their evaluation of the model.

The authors would like to thank Dr Roger Wood and Dr Andrew Damant (FSA) for their encouragement during this work and for providing the GMO data for statistical analysis. We also wish to thank Dr Eric Wyse and Dr Stephen de Mora (IAEA) for their useful discussions and for the use of the chromium data (IAEA 405).

Our thanks to Dr Ian Davies for the discussions on the model and, in particular, his contribution to the assessment of LCVs.

Our thanks also to the QUASIMEME team and to the many QUASIMEME participants who have provided endless datasets on which to test the model.

10. REFERENCES

- Analytical Methods Committee. 1989a. *The Analyst*, **114**(1992), 1693-1697.
- Analytical Methods Committee. 1989b. *The Analyst*, **114**(1992), 1699-1702.
- AMC. 2001a. Technical Briefs No 4. Representing Data Distributions with Kernel Density Estimates April. Royal Society of Chemistry www.rsc.org/lap/rsccom/amc_index.htm
- AMC. 2001b. Technical Briefs. No6 Robust Statistics: a method of coping with outliers April. Royal Society of Chemistry www.rsc.org/lap/rsccom/amc_index.htm
- Asmund, G. 2002. Private Communication. Based on a presentation made at the QUASIMEME Conference, Barcelona October 2002. QUASIMEME Project Office, FRS Marine Laboratory, 375 Victoria Road, Aberdeen AB11 9DB.
- de Boer, J., Oehme, M., Smith, K. and Wells, D.E. 2000. Toxaphene in standard solutions and cleaned biota extracts - results from the first QUASIMEME interlaboratory studies. *Chemosphere*, **41**, 493-497
- de Boer, J. and Wells, D.E. 2002. BSEF/QUASIMEME Interlaboratory Study On Brominated Flame Retardants Exercise 524 June. QUASIMEME Project Office, FRS Marine Laboratory POB 101, Victoria Rd Aberdeen, AB11 9DB.
- Cofino, W.P. and Wells D.E. 1994. Design and Evaluation of the QUASIMEME Inter-Laboratory Performance Studies: A Test Case for Robust Statistics. *Mar. Poll. Bull.*, **29**, 149-158.
- Cofino, W.P., Wells, D.E., Ariese, F., van Stokkum, I.H.M., Wengener, J.W. and Peerboom, R. 2000. A new model for the inference of population characteristics from experimental data using uncertainties. *J. Chemometrics and Intelligent Laboratory Systems*, **53**, 37-55.
- Cofino, W.P., van Stokkum, I.H.M., van Steenwijk, J. and Wells, D.E. 2004. A new model for the inference of population characteristics from experimental data using uncertainties. Part II. Application to censored datasets. *Anal. Chim. Acta* (in press)
- Dorsey, E.N. 1944. The Velocity of Light. *Trans. AM. Philosophical Soc.* **34**, p+1, 1-100 (Table 22).
- Gilbert, O.R. 1987. *Statistical Methods for Environmental Pollution Monitoring*. John Wiley and Sons. N.Y., ISBN 0-471-28878-0
- Gilliom, R.J. and Helsel, D.R. 1986. Estimation of distributional parameters for censored trace level water quality data. 1. Estimation techniques. *Water Resour. Res.*, **22**, 135-146.
- Helsel, D.R. and Gilliom, R.J. 1986. Estimation of distributional parameters for censored trace level water quality data. 2. Verification and applications. *Water Resour. Res.*, **22**, 147-155.
- Helsel, D.R. and Cohn, T.A. 1988. Estimation of descriptive statistics for multiply censored water quality data. *Water Resour. Res.*, **24**, 1997-2004.

- Huber, P.J. 1972. Robust Statistics: A Review. *Annals of Mathematical Statistics*, **43**, 1041-1067.
- Hoaglen, D.C., Mosteller, F. and Tukey, J.W. 1983. Understanding Robust and Exploratory Data Analysis. Wiley.
- ISO. 1994. ISO 5725 Accuracy (trueness and precision) of measurement methods and results. (Parts 1-5).: International Standard Organisation Geneva, 1994.
- ISO. 1997. Guide 43-1 Proficiency Testing by interlaboratory comparisons. International Standard Organisation Geneva.
- Kuttatharmmakul, S., Massart, D.L., Coomans, D. and Smeyers-Verbeke. 2001. Comparison of methods for the estimation of statistical parameters of censored data. *J. Anal. Chim. Acta*, **441**, 215–229
- Lischer, P. 1987. Robuste Ringversuchsauswertung. *Lebensmittel-Technologie.*, **20**, pp. 167–172.
- Lothian, P.J. and Thompson, M. 2002. Bump hunting for the proficiency test er- searching for multimodality. *The Analyst*, **127**, 1359-1364.
- Mavrodineanuv, R. 1971. NIST. Certification for Transmittance value for a filter.
- Montville, D. and Voigtman, E. 2003. Statistical properties of limit of detection test statistics. *Talanta*, **59**(3), 461- 476.
- Silverman, B.W. 1986. Density Estimation for statistics and data analysis. Monographs on Statistics and Applied Probability No 26, Chapman Hall.
- Thompson, M. and Wood, R. 1993. The international harmonized protocol for the proficiency testing of (chemical) analytical laboratories. *Pure and Applied Chemistry*, **65**(9), pp2123-2144.
- Wand, M.P. and Jones, M.C. 1995. Kernel Monographs on Statistics and Applied Probability No 60 Chapman Hall.
- Wells, D.E. and Cofino, W.P. 1997. The Assessment of the QUASIMEME Laboratory Performance Studies Data: Techniques and Approach. *Mar. Poll. Bull.*, **35**, 10-27.
- Wells, D.E. and Hess, P. 1999. Methods for the determination and evaluation of chlorinated biphenyl's (CBs) in environmental matrices. Chapter 6, pp 239-286 in Sample Handling and Trace Analysis of Pollutants, Techniques, Applications and Quality Assurance ed. D. Barcelo, Elsevier.
- Youden, W.J. and Steiner, E.H. 1975. Statistical Manual of the AOAC, Association of Official Analytical Chemists.

ANNEX 1

THE INPUT AND OUTPUT OF THE COFINO MODEL

The Data Input for the Cofino Model

There are different configurations of data input possible for the basic Cofino model. As a minimum, the model requires a row vector of single observations for each determinand in each separate matrix. For single measurements the model provides an evaluation based on the Normal Distribution Assumptions (NDA) and the BandWidth Estimator (BWE) model, calculated from the values provided.

Where these values are the means of measurements then the standard deviation (s.d.) can also be provided. Replicate data can also be provided as a matrix with the observations as columns and the replicates as rows.

Two options are possible. Firstly, with two row vectors, one for the mean and one for the s.d. Secondly, with a matrix as observations.

The model calculation only occurs with the number of observations (NObs) > 4 for the NDA model and to NObs >10 for the BWE model. Only the NDA model can be used between NObs = 5 and =10.

Additional input may also be provided for the laboratory identifiers (default = 1,2,3,...)

The basic model can be automated with information on file locations, and parameters to calculate z scores for assessment purposed.

The Information Output

The summary output provides the mean, s.d. and percentage of data that comprise the mean for the first 7 modes of the data using both the NDA and the BWE models. In addition the cumulative sum of the percentage of data is also provided. This output gives a full quantitative description of the dataset. Where replicates are used the BWE output is replaced by that of the replicates.

Robust statistics [Lischer 1987 and AMC 1989a & b] are provided in parallel to Cofino model statistics, along with the normal statistics, primarily for comparison.

Other information that is generated includes the Kolmogorov-Smirnov test for normality and the Students t value and an indicator for bimodality. The Students t value is obtained from the means of s.d's of the first and second modes.

These output statistics are use in the evaluation of the dataset.

An example of the output is given in Table 1.

TABLE 1

An example of the primary statistical output from the Cofino model, robust and normal statistics for 3 PAHs. Data are exported from the MATLAB program into an Excel spreadsheet.¹³ These data are used to create the summary statistics Tables in the participants' reports.

	Sample Determinand		QPH029BT Acenaphthylene	QPH029BT 2-Methylphenanthrene	QPH030BT Naphthalene
	Nobs	all	14	7	18
	Assigned	final	1.45	2.96	2.55
	NOBs	Less Than	3.00	0.00	1.00
NDA	PMF1	Mean	1.65	3.03	3.56
BWE	PMF1	Mean	1.45	2.96	2.55
Robust	Lischer	Mean	2.58	3.85	5.36
Robust	AMC	Mean	2.86	3.72	6.12
NDA	PMF2	Exp val	3.21	17.21	12.59
BWE	PMF2	Exp val	1.67	17.30	6.35
NDA	PMF1	s.d.	1.52	2.64	3.77
BWE	PMF1	s.d.	0.83	1.85	2.11
Lischer		s.d.	2.38	2.49	5.34
AMC		s.d.	2.75	2.57	6.02
NDA	PMF2	s.d.	2.12	2.62	6.92
BWE	PMF2	s.d.	1.25	1.54	3.02
Students		t Value	0.26	8.20	2.18
NDA	PMF1	% Data	67.74	78.54	67.75
NDA	PMF2	% Data	10.61	14.33	16.86
NDA		%-cum.	78.34	92.87	84.61
BWE	PMF1	% Data	53.98	70.77	52.52
BWE	PMF2	% Data	17.72	14.29	20.98
BWE		%-cum	71.70	85.05	73.50
Kolmogorov	Smirnov	Normality	1	1	1
Cofino	s.d.	Within			
AMC	s.d.	Within			
Normal		Mean	39.6	5.1	10.5
Normal	Between	s.d.	115.9	5.6	19.5
Normal	Within	s.d.			
	Median		1.8	3.6	3.2
Proportional	Error		12.5	12.5	12.5
Constant	Error		0.5	5.0	1.0
Indicative			1.0	1.0	1.0
BWE	CV%		67.8	83.0	73.3
AV	Uncert %		20.9	28.1	27.7
Target	CV%		28.2	97.7	35.0
BWE			0.6	1.5	1.5
Percent			54.5	71.4	41.2
BWE	Nobs		6.0	5.0	5.0
Mean	Percentile		36.0	43.0	53.0
KDE	H opt		1.0	1.6	2.4

¹³ Actual data output is transposed in this paper to aid formatting.

Graphical Representation of the Data

In addition to the tabular summary statistics of the data for the Cofino model or for the robust statistics it is informative to provide a graphical representation of the distribution of the data.

Four types of plots have been selected to provide a graphical representation of different aspects of the data.

These cover:

- the distribution of the data
- the extent of overlap and modality of the data
- the distribution of real and left censored values
- the level of agreement of data with the assigned value

Four data sets have been selected to provide examples of some of the key characteristics of different data sets and how these are displayed in the plots. The four sets of data are:

	Matrix	Name	NObs Real	NObs LCV	Assigned Value	Coefficient of variation
Fig 2	QNU118 SW	Nitrite in Seawater	46	17	0.023 umol/L	64
Fig 3	QOR070 BT	HCB in Biota	24	9	0.064 ug/pg	35
Fig 4	QTM065 MS	Cadmium in Marine Sediment	52	0	11.042 ug/kg	15
Fig 5	QOR077 BT	"-HCH in Biota	12	13	0.026 ug/kg	54

Ranked Overview

In most studies, participants report a single value that is the best estimate of the concentration of each determinand. A plot of these data ranked by magnitude with error bars of $\pm 2 \text{ s.d.}_{\text{NDA}}$ or $\pm 2 * \text{ s.d.}_{\text{bwe}}$, depending on calculation, provide an overview of the distributions of the data and the degree of overlap of the values.

Where the data provided as means and s.d. or replicate values, the mean value and s.d. are plotted. The minimum range of values plotted is either zero or the mean – 3s.d., whichever is the smaller. The maximum range is limited to the mean +10 s.d. This allows a reasonable detail to be plotted.

Numerical data are plotted in blue as means $\pm 2\text{s.d.}$. Where the value exceeds the mean +10 s.d., a blue dotted line is plotted.

The Left Censored Values (LCVs)¹⁴ are plotted as red circles connected to zero by a solid or dotted line. A dotted line indicates that the LCV is > mean of the numerical data and is excluded from the estimation of the assigned value (AV). This is the only limitation on the data. All other data are included in the assessment. The exclusion of these data while establishing the AV is necessary since large LCVs do not contribute to the information on the population characteristics. Because the LCVs have a rectangular distribution function (0 to LCV) they can also overlap with the real value between 0 and the LCV. There is, in turn, means that the large LCVs link with the extreme positive numerical values and falsely elevate the mean (Cofino *et al* 2004).

¹⁴ Left Censored Values is the appropriate nomenclature for “less than” values

The LCVs are ranked below the lowest real value in reverse order of magnitude. The range of the LCV is given, and plotted from zero to the reported LOD in place of the error bars calculated for the positive values.

The Cofino model mean is given as a solid line. This is the best estimate of the mean of these measurements.

The overall *within laboratory* standard deviation (s.d. _{within}) is estimated using the BWE and the NDA of the Cofino model, (Cofino *et al.*, 2000).

$$\text{s.d.}_{\text{within}} = c_i \cdot \text{MAD where } c_i \text{ is the coefficient derived from the model}$$

The MAD is the median absolute deviation given by

$$\text{Median}(\chi_i - \text{median}(\chi))$$

The coefficient c_i has a value of 1.168 when the MAD is constructed on the whole data.

The bandwidth estimator (BWE) selects a *heart-cut* of the data about the mode, ie values with the highest level of overlap, to construct the s.d. In this case the value of the coefficient, c , is dependent on the percentage of the whole dataset that is used. Details of the model and the coefficients of this correlation are given in the report. In most cases the BWE provides more information on the structure of the data. Details and examples are given in the Cofino model handbook (Wells *et al.*, 2004).

Extreme positive values reported are shown on the Ranked Overview as dotted lines (blue), in order to maintain a scale to provide sufficient detail.

These data are included in the data assessment and the construction of the assigned value.

Summed Probability Density Function¹⁵

The graphical representation of the summed pdfs provides a profile of the distribution of the data. Using the BWE the profile gives useful information on the presence of bi or multimodal data. The mode of the first seven PMFs are also added to the trace, with PMF, indicating the main mode of the data, which for most data sets is equivalent to the assigned value.

LCV data that are greater than the mean of the numerical data are not included in the Cofino BWE model for the estimation of the assigned value and, therefore, are not included in this plot.

Kilt Plot (Overlap Matrix)

The overlap integral is derived from the Cofino model. The matrix of the overlap integral of each value with all other values is represented in a 2 dimensional plot. The observations are ordered according to their magnitude from low to high concentrations and the degree with which they overlap with neighbouring values is denoted by the colour of the matrix. When

¹⁵ In the Cofino model each observation is not regarded as a value but is described by a probability density function (PDF). Analogous to a wavefunction the observation measurement function (OMF) is the square root of the PDF. From this the population measurement function (PMF) can be constructed by a linear combination of the OMF's related to each observation. The normalized, square PMF is the resultant summed PDF. The detailed discussion of this model is given elsewhere (Cofino *et al.* 2000, Cofino *et al.* 2004, Wells *et al.* 2004).

present LCVs are plotted together from the left corner of the plot. The colour reflects directly the magnitude of the integral. Areas of the map coloured white indicate complete overlap (agreement) for the observations concerned and black indicates no overlap.

The 2 dimensional plot of the overlap matrix provides a diagonal line of white square on every plot since the overlap between a basisfunction with itself is plotted along this axis and these overlaps are, by definition, equal to 1.

Such a plot provides an immediate impression of the level of agreement between the laboratories and of the distribution of the data. The term “Kilt” plot reflects the chequered, tartan, nature of the presentation.

LCVs that are greater than the mean of the numerical values are not included in the Kilt Plot. The lighter the total figure the better the overall agreement.

Z Score Plots

The Z score plots are a ranked representation of the data as a bargraph, normalised to the assigned value and the target s.d., such that

$$z \text{ score } (Z) = (\chi_i - \mathbf{x}) / \text{s.d.}$$

Where χ_i is the individual value, \mathbf{x} is the assigned value and s.d. is the target s.d. In some schemes the target s.d. is derived from the dataset, while in others it is set as an external standard of achievement. In the former, 95% of the data would achieve values of $|Z| < 2$ when the dataset is approximately normal. QUASIMEME applied external standards for the performance requirements set by the Scientific Assessment Group.

The ISO 43 (ISO 1997) for laboratory performance study assessments provides three levels of evaluation. A z score of $|Z| < 2$ is classified satisfactory, $2 < |Z| < 3$ is questionable and $|Z| > 3$ is unsatisfactory.

QUASIMEME has added a further category of $|Z| > 6$ for assessment purposes. These are extreme values and may be regarded as not belonging to the population. They may be caused by the use of incorrect units, dilution /concentration or calculation errors rather than intrinsic errors in analytical methodology. Factors of 2,5,10 or 1000 are common for these types of extreme values.

In addition to the regular assessment for real values, QUASIMEME has introduced a further two categories for LCVs.

Until now LCVs have only been recorded and not evaluated. However, now that the model includes these values in the overall data distribution and calculation of the Cofino mean it is important to provide a more formal assessment of these values. This assessment is especially needed where the number of LCVs constitutes a significant fraction of the dataset.

Although LCVs $>$ mean of the numerical values are excluded from the estimation of the assigned value all LCVs are assessed. The basis of the assessment is to confirm (or otherwise) that the range covered by the LCV is consistent with the assigned value. A LCV (e.g. < 0.1) considerably smaller than the assigned value (e.g. 1) would imply that the method should have detected the determinand. This is not consistent with the assigned value.

Likewise a LCV which was much greater than the assigned value would imply that the method used was not sufficiently sensitive to detect the determinand, compared with other results.

The upper and lower boundaries selected are related to the target performance set by QUASIMEME to provide a common level of assessment. These boundaries are equivalent to $|Z| < 3$, where Z is the equivalent magnitude to the Z score and derived using the same formulae, as given above.

The actual value included in the assessment is 50% of the LCV. The rationale for this is simply taken from the distribution characteristics of the LCV, which is regarded as a rectangular probability distribution.

When this is compared to the normal distribution characteristics of a numerical value, it is clear that the actual reported value is at the 50 percentile of the probability distribution (Figure 1). In both cases the centre of a symmetrical distribution is used.

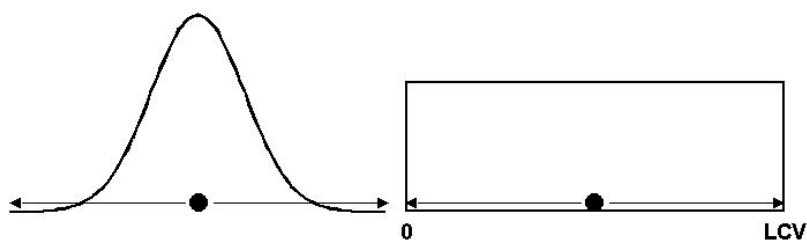


Figure 1. The normal probability distribution of a real value and the rectangular probability distribution of a Left Censored Value (LCV). The actual LCV reported in the upper boundary i.e. “less than” of the distribution that ranges from 0 to the LCV.

The two categories that have been assigned to the LCVs are:

- [1] **onsistent with the assigned value.** LCVs which fall into the category ‘consistent’ have a value $LCV/2$ which is $< |Z| \leq 3$.
- [2] **nconsistent with the assigned value.** LCVs which fall into the category ‘inconsistent’ have a value $LCV/2$ which is $> |Z| \leq 3$.

The categories consistent and inconsistent are separate from **satisfactory** and **unsatisfactory** which are reserved for the assessment of numerical data.

LCVs that are assessed as **Consistent with the assigned value** are shown on the z score plot with a red circular symbol.

Examples of the Plots

Nitrite in Seawater (QNU118SW)

This test material was natural seawater with very low levels of nutrients (Nitrite 0.023 $\mu\text{mol/l}$). Of the 63 observations, 17 were reported as LCVs.

The summed probability density function shows a number of steps in the distribution due to the integration of the rectangular pdfs from the LCVs and the normal pdf from the real values.

The Kilt Plot shows a large central square of numerical data in good agreement and the 12 LCV that has values consistent with the assigned value. These values are also shown on the ranked Z score plots.

The dark bands on the Kilt Plot relate to the six extreme positive observations that have almost no overlap with most of the main body of data. Note that the 5 LCVs that are shown as dotted red lines are greater than the mean of the numerical data.

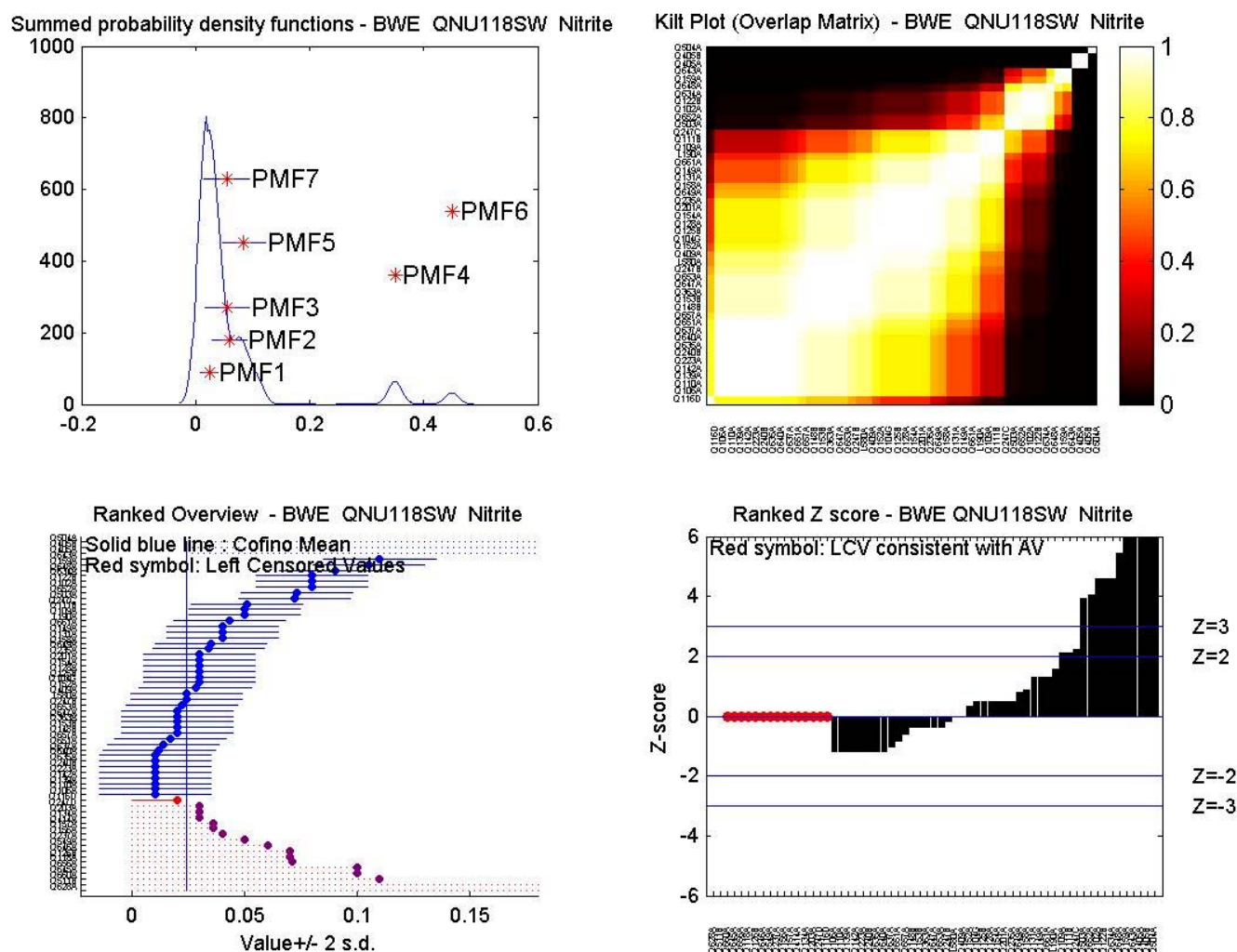


Figure 2. Nitrite in seawater (QNU118SW). Graphical summary plots

These values are excluded from the Cofino model and do not occur either in the probability density function trace or in the Kilt Plot.

Three of these extreme positive values are shown as dotted blue lines at the top of the ranked overview. There is a group of 8 observations having the same value (0.01 umol/L) which is on the lower limit of detection. Of the 17 LCVs, 12 are consistent with the assigned value, while the remaining 5 LCVs are both $> \text{mean} + 2\text{s.d.}$ and greater than the equivalent of $Z = 3$.

Hexachlorobenzene (HCB) in Biota (QORO76BT)

Of the 33 observations, 24 were real values and 9 were LCVs. Again the summed probability density functions are stepped due to the influence of the rectangular PDFs from the LCVs. The 5 *modes* at higher concentrations are created by the five extreme positive values, with some overlap from 2 other values around 0.2-0.25 ug/kg.

These extreme positive values are seen as single white or white/orange squares in the black bank of the Kilt Plot. These data have no overlap with the other values. There is relatively good agreement between 15 values around the main mode between 0.060-0.13 ug/kg.

The 5 LCVs consistent with the assigned value 0.068 ug/kg that also overlap with this group of 15 real values. This can be seen both in the Ranked Overview and the Kilt Plot.

Four of the LCVs are inconsistent with the assigned value. These values are off scale on the Ranked Overview plot and depicted by dotted red lines.

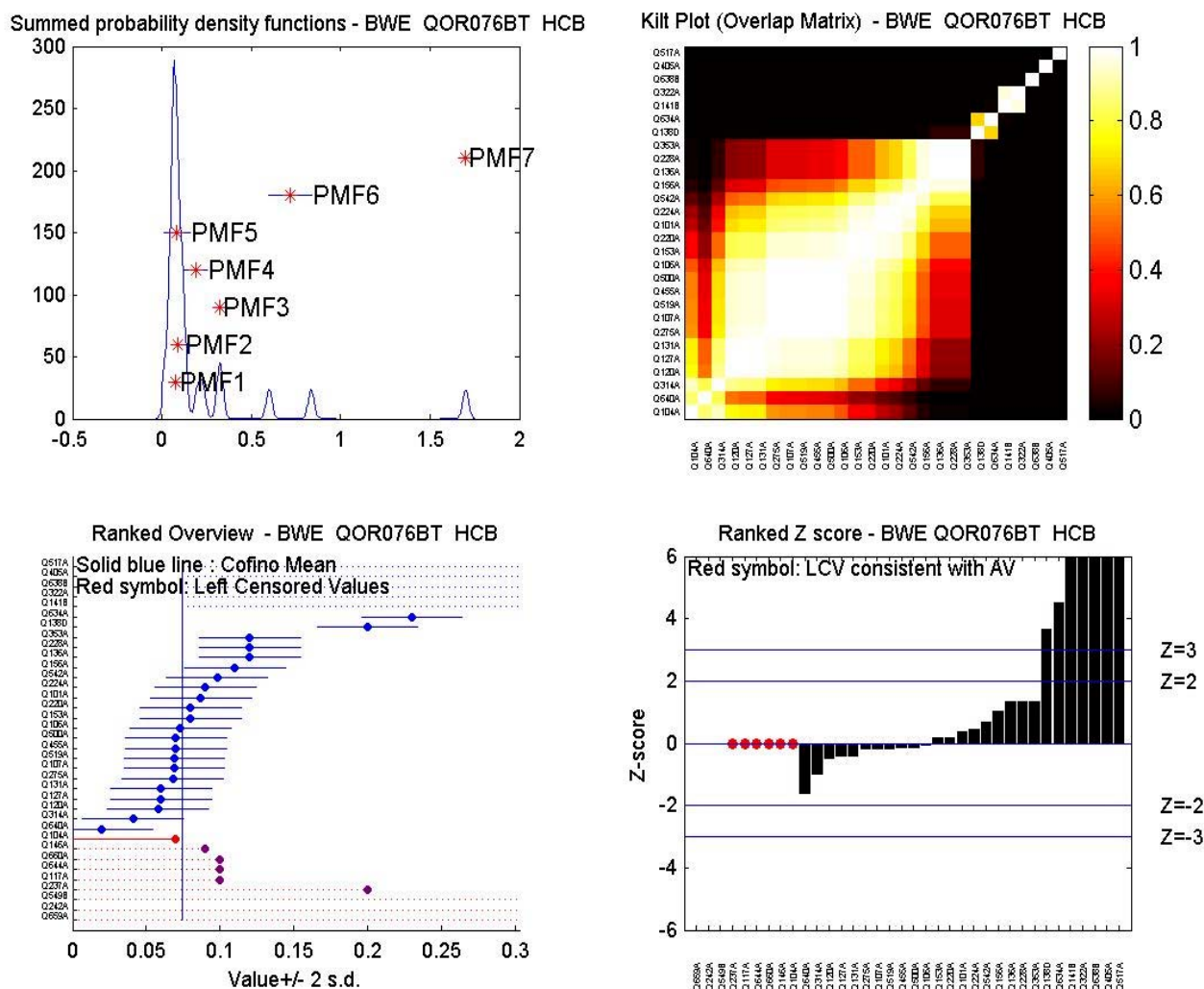


Figure 3. Hexachlorobenzene (HCB) in Biota (QOR076BT).Graphical summary plots

Cadmium in Marine Sediment (QTM065MS)

All 52 observations were reported as numerical values. The most striking feature of these plots is that the data are clearly modal with the first, main mode around the assigned value, 11.0 ug/kg and a second mode of which there are two groups of data. With the present distribution of the data the BWE used in the Cofino model would need to be reduced in order to separate the second mode into its two components. However, separating the data into two modes is sufficient. Of the 10 observations in the second mode, 6 are from new laboratories and one other has a trainee as the analyst. The factors separating the modes are either 5 or 1000, the latter due to incorrect reporting units.

Those laboratories that comprise the first mode show very good agreement.

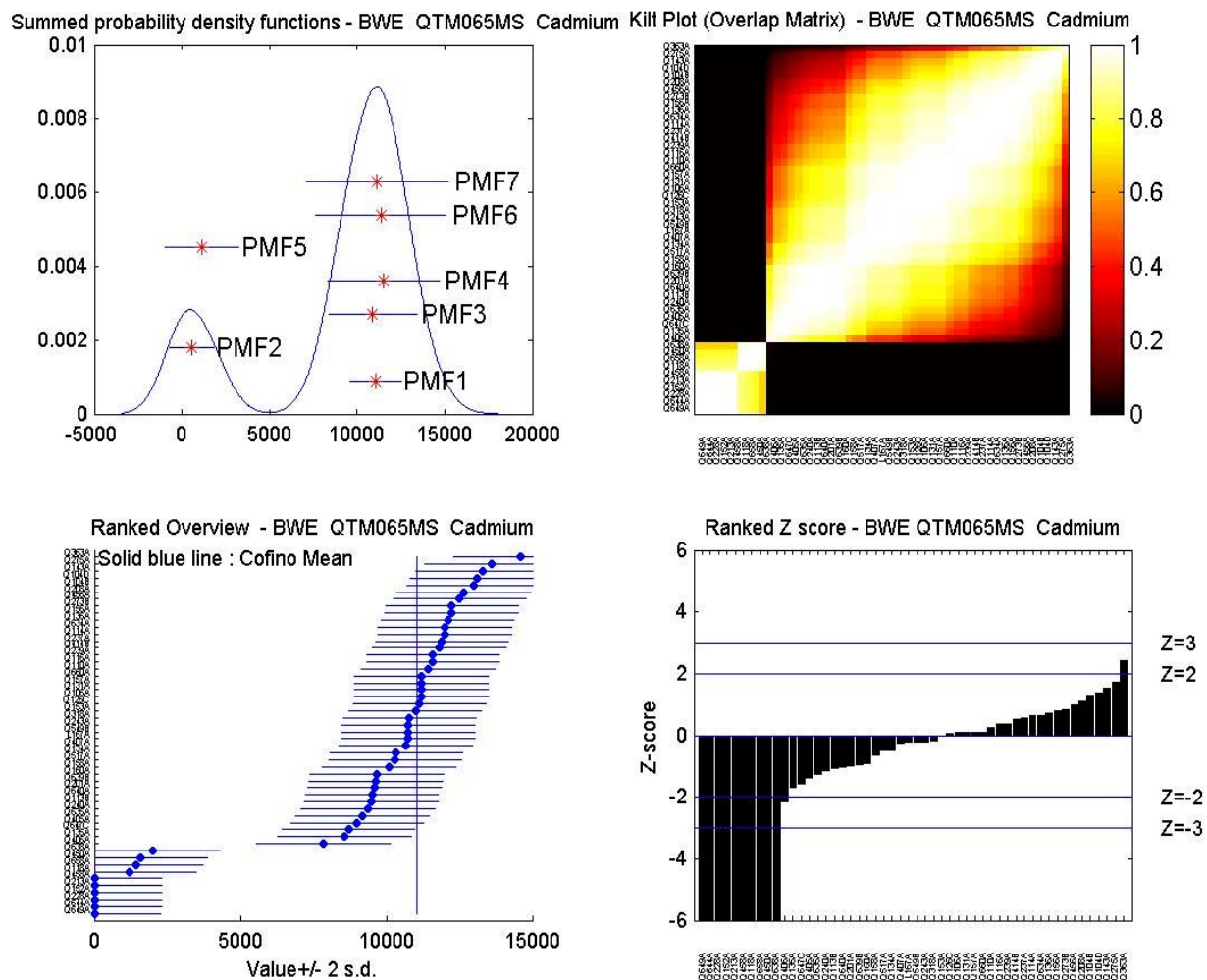


Figure 4. Cadmium in Marine Sediment (QTM065MS). Graphical summary plots

"-HCH in Biota (QOR077BT)

This is a relatively small dataset with 25 observations, 12 of which are numerical and the remaining 13 are LCVs. The Kilt Plot provides a clear view of the distribution of the data which is primarily bimodal between the 8 numerical data which are satisfactory ($|Z| < 2$) and the second mode with 7 LCVs which are consistent with the assigned value (0.026 ug/kg). There are 6 LCVs which are not consistent with the assigned value ($> Z = 3$ and $> \text{mean} + 2\text{s.d.}$). There are also two extreme positive values that appear as blue dotted lines at the top of the Ranked Overview and as two single squares in the upper right of the Kilt Plot.

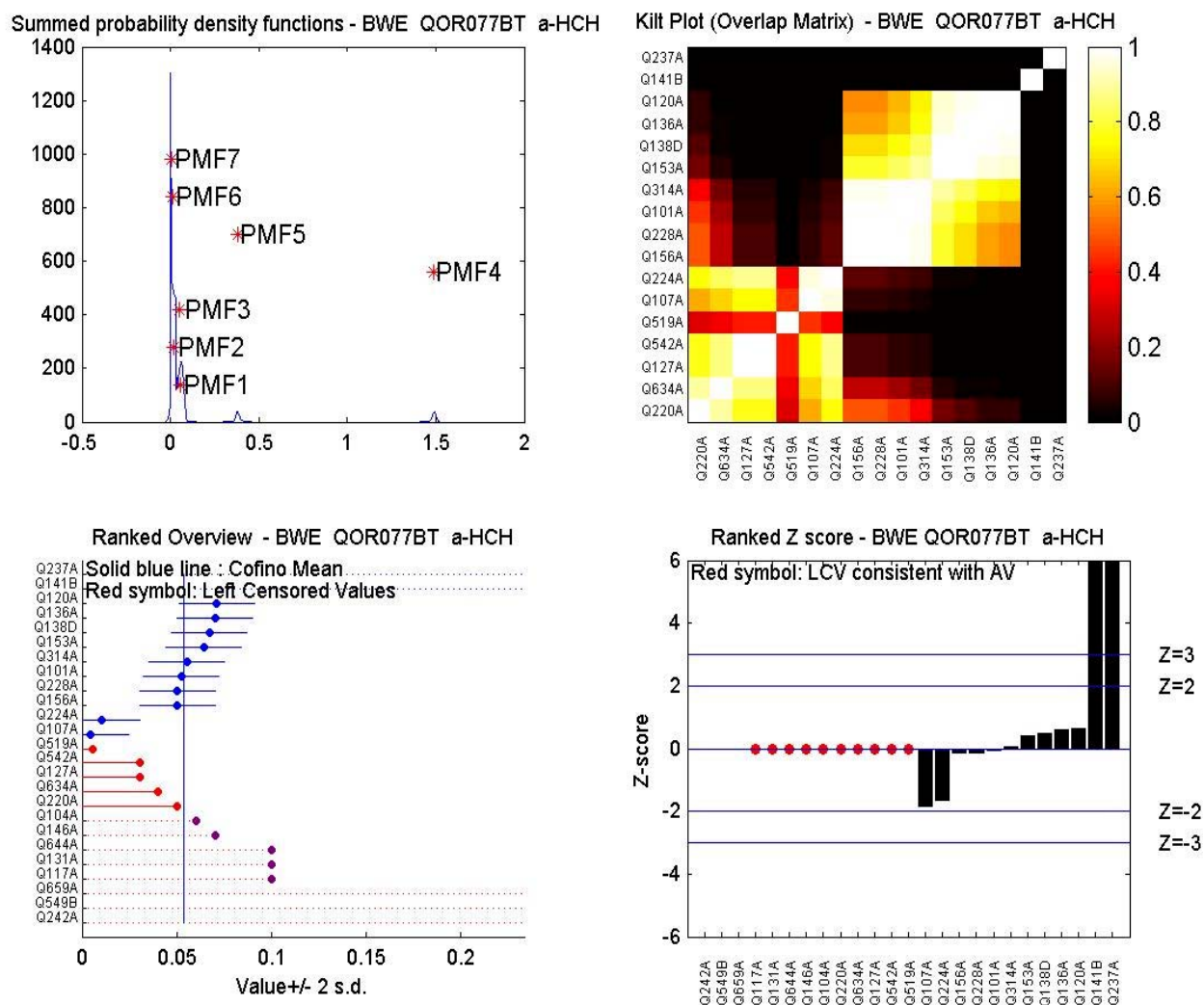


Figure 5. "-HCH in Biota (QOR077BT). Graphical Summary Plots

Fisheries Research Services is an agency of the Scottish Executive

FRS MARINE LABORATORY PO Box 101 375 VICTORIA ROAD ABERDEEN AB11 9DB UK

TEL +44 (0)1224 876544 FAX +44 (0)1224 295511

FRS FRESHWATER LABORATORY FASKALLY PITLOCHRY PERTSHIRE PH16 5LB UK

TEL +44 (0)1796 472060 FAX +44 (0)1796 473523

enquiries@marlab.ac.uk <http://www.marlaboratory.ac.uk>