## Supporting Information

# Novel Method for Calculating a Nonsubjective Informative Prior for a Bayesian Model in Toxicology Screening: A Theoretical Framework

Michael Woldegebriel*

Analytical Chemistry, Van 't Hoff Institute For Molecular Sciences, University of Amsterdam

P.O. Box 94720,1090 GE Amsterdam, The Netherlands

E-mail: M.T.Woldegebriel@uva.nl

# Contents

## SIMULATION OF ANALYTICAL METHODS

In order to demonstrate the proposed formulation, simulated analytical method (reflecting analytical methods used in toxicology screening), and data adopted from Netherlands Food and Consumer Product Safety Authority (NVWA) was used.

### Simulation of Analytical Method Performance Boundaries

The selected analytical methods (depicted in Fig.2 in the manuscript) were assigned a specific value reflecting their performance boundaries (by using random number generator). This step is representative of the spike-in approach that is normally required to experimentally estimate the unknown parameters. For the random number generator, a reasonable boundary was assigned; first, taking in to account the data obtained from NVWA has already made an exhaustive use of existing analytical methods, a set of values containing the minimum observed CL (in mg/kg) for all 153 pesticides was created. The minimum and maximum possible values from this set were then assigned as the minimum and maximum boundaries for the random number generator, respectively. Fig.S1 depicts a graphical representation of the analytical method performance boundaries assigned.

### Grid Representation of Spike-in Biological Samples

As discussed in the theory section of the manuscript, in order to estimate analytical method parameters, biological samples spiked-in with known CL of COI are required. In order to represent this experimental step with simulation, a numerical grid starting from 0 up to CL of saturation (computationally; where the parameter DSL is obtained) was created. Each CL that would have been used for spike-in was then represented numerically by selecting one of the values from the numerical grid iteratively in ascending order. For each iterative step, the selected

CL was then matched with the simulated analytical method boundaries using a Bayesian model described in the following section.

## ESTIMATION OF 'COI' OBSERVATION PROBABILITY DISTRIBUTION

In order to estimate analytical method parameters, the probability distribution of COI observation versus CL is required (as illustrated on Fig.3 in the manuscript). With the simulated analytical method performance boundaries, and numerical grid representation of spike-in CL, an iterative process that applies the Bayesian formulation presented on Eq.(6-9) was performed. The aim of this step is to estimate the analytical parameters (MDL, MIL, C*, and DSL). Therefore, the iteration was run until $P(H_2|D)$ converges to 1 (referring to maximum probability), in which at that point the CL from the numerical grid at that iterative step was adopted as DSL. The iterative step explained can be formulated as follows:

$$\begin{cases} \text{for } g = 1 - G \\ \quad \text{while } P(H_2|D, g) < 1 \\ \quad\quad \text{continue} \\ \quad \text{else} \\ \quad\quad \text{break} \\ \quad\quad DSL = CL_g \\ \quad \text{end} \\ \text{end} \end{cases} \tag{S1}$$

On Eq.(S1), g refers to the index of a given CL from the numerical grid, the data (D) at every step of the iteration represents the chromatogram obtained by spiking a given biological sample with $CL_g$ of COI. As indicated on Eq.(9), a marginalization step to account for all possible combinations of considered analytical methods is necessary. In order to represent real case scenario, where one has no a priori information about the methods in question, the probabilities, $P(I_i|H_2)$ and $P(S_s|H_2)$ in the formulation were assigned a flat probability distribution. Here it should be noted that the probabilities are basically indicating how probable it

is for the instrument ($I_i$) and software ($S_s$) in question are capable to support the hypothesis ($H_2$). Similarly, the probability for the likelihood, $p(D|I_i, S_s, H_2)$, that would in a normal case require feature extraction algorithm, was assigned a discrete probability value based on function f as follows:

$$f = \begin{cases} \text{if } CL_g < I_i \text{and } S_s & 0 \\ \text{if } CL_g \geq I_i \; and \; CL_g < S_s & 0.5 \\ \text{if } CL_g < I_i \; and \; CL_g \geq S_s & 0.5 \\ \text{if } Cl_g \geq I_i \text{ and } S_s & 1 \end{cases} \quad \text{(S2)}$$

On Eq.(S2), $I_i$ and $S_j$, represents the simulated performance boundary values assigned by the random number generator, for instrument and software, respectively. $CL_g$ refers to the randomly selected unique CL value at iteration g. The purpose of function f is to evaluate if any given combination of a software ($S_s$) and instrument ($I_i$) with known performance boundaries are capable of identifying COI at a given CL.

## Extraction of Analytical Method Parameters from the Simulated System

Given the combination of all analytical methods considered, the first CL that obtains a posterior probability above 0.5 was used as an indication of the optimal CL necessary for positive identification of COI (probability of 0.5 taken as a logical boundary to identify when hypothesis $H_2$ is true). Therefore, by applying the same parameter extraction as depicted on Fig.3 in the manuscript, a value of 0.4119 mg/kg was adopted as MDL. Similarly, CL of 0.2074 mg/kg, which obtained the first posterior probability above 0, and CL of 0.8201 mg/kg which obtained the first posterior probability of 1, was assigned as MIL and DSL, respectively. For the case of C*, CL of 0.00046 mg/kg; computed as the distance between the CL needed to obtain MIL and the CL needed to cause the first deflection that CL, was assigned.

## K-means Clustering

As discussed in the manuscript, after obtaining the analytical method parameters, the next step necessary would be to create CL grid (CLG) representing all analytically meaningful CL. In a follow up step, the CLG will be adjusted by a large dataset consisting of representative CL of COI by approaching it as a clustering problem. The method proposed for this classification (translation) of observed continuous CL (CCL) in to discrete CL (DCL) step is K-means clustering. Here it should be noted that the cluttering step can be interpreted as adjusting the analytically meaningful CLG using large dataset, or can also be interpreted as discretizing the observed large dataset in to analytically meaningful clusters using CLG. Either ways, both interpretations reflect on the same goal. Below are the steps to be followed during the classification (clustering) process:

I. Distance between each CL in the observed data and all the CL in CLG will be calculated.

II. Based on the calculated distance, all the random samples within the large dataset will be allocated to the nearest CL value in the CLG.

III. A new centroid is determined for each cluster by calculating the mean of samples in each cluster.

IV. Distance between each CL in the large dataset and the new centroid values will be calculated

V. Each sample is once again assigned to the nearest centroid.

VI. Step III-V is repeated until convergence (the centroid values does not change indicating the variance within cluster is less than between clusters).

Once the clustering step has converged, the final centroids of each cluster are accepted as the DCL representation of the CCL of the large dataset.

## PROBABILISTIC BOOTSTRAP SAMPLING

As discussed in the K-means cluttering section, one of the steps implemented in the method proposed is clustering. In this step, as mentioned, a large dataset consisting of representative CL of COI is classified to a specific CL by implementing CLG (created with method parameters) as an initial centroid. In this clustering step, in addition to using the observed large dataset, a bootstrap sampling is also conducted. This way, unobserved case scenarios in terms of CL can be contemplated.

For the sampling process, two basic steps are followed. On the first step, a set consisting of elements representing index of each value in the large dataset is created. The elements in this set are then randomly selected with replacement. After the appropriate randomly selected index elements has been obtained, in contrary to the traditional bootstrap sampling approach where the same elements from the original dataset are selected based on the randomly selected indexes, here a probabilistic approach is followed. For a given randomly selected CL referred by a given randomly selected index, the possibility for the index to represent any other CL within the vicinity of the CL in question is contemplated. This selection process applies a normal probability distribution, taking the CL in question (in which the index is referring to) as the mean of the probability distribution, and $\frac{C*}{2}$ as its standard deviation. The probability of a randomly drawn new CL can be expressed as follows:

$$P(CL_n|CL_x) = \frac{1}{\left(\frac{C*}{2}\right)\sqrt{(2\pi)}} e^{-\frac{1}{2}\left(\frac{CL_n - CL_x}{\left(\frac{C*}{2}\right)}\right)^2} \quad \text{(S3)}$$

On Eq.(S3) $P(CL_n|CL_x)$ refers to the probability of the newly selected CL ($CL_n$) given the CL in question ($CL_x$). The reason for choosing such a parameter for the standard deviation is based on the defined analytical method parameters, in which C* has already been identified as the minimum possible CL needed to create a variation in system response. Therefore, with this understanding, as long as the variation between $CL_x$ and $CL_n$ does not exceed C* (95% of the cases remain within$CL_x \pm C^*$), it can be considered as the new bootstrap sample. This way, exhaustive case scenarios can be contemplated while still constraining the approach by the analytical method parameters.

## VALIDATION OF THE PROPOSED METHOD

### Validation One

To validate the method proposed, reference data that can be used as the ground truth, reflecting all possible CL to be encountered in a laboratory setting, is compulsory. All statistical approaches are based on the assumption that there is a hypothetical population of samples from which the observed data are drawn. In reality, since it is not possible to obtain such a data, a validation step which adopts the data obtained from Netherlands Food and Consumer Product Safety Authority (NVWA), consisting of 4896 screening results of 153 pesticides from 2151 fruits and vegetable samples obtained between January and July 2014, as the ground truth was formulated. The general idea behind this validation method is to show that, if a set of randomly drawn samples from the NVWA data (now adopted as the ground truth) contains sufficient information regarding the CL of COI, the final $P(H_0)$ obtained for the random samples using the method proposed should converge towards the $P(H_0)$ of the NVWA data itself. Here it should be noted that, the analytical method boundaries simulated for demonstrative purposes (Fig.S1) are

once again adopted for this validation step. Therefore, taking three pesticides as an example (Imazalil, Thiabendazole, and Ortho-phenylphenol), 2151 results were extracted for each. Next, from the extracted dataset, only 30% of the uniquely observed CL for each pesticide was randomly drawn and adopted in the computation (Fig.S2). In this way, the worst-case scenario can be contemplated and the robustness of the method can be tested. The 30% randomly drawn samples represent the large dataset of random samples required to apply the proposed method. Given the randomly drawn samples, the computational steps explained in the theory section (estimation of the analytical methods, probabilistic bootstrap sampling, clustering approach, and random draw approach) for each pesticide were followed. Fig.S3A-C depicts the clusters generated taking few bootstrap samples as an example. The final $P(H_0)$ obtained as the maximum likelihood estimate of the posterior probability distribution (depicted on Fig.S3D-F) for Imazalil, Thiabendazole, and Ortho-phenylphenol were 0.126, 0.171, and 0.180, respectively. Similarly, the $P(H_0)$ for the extracted CL from NVWA dataset for all pesticides (adopted as the ground truth) was also computed using direct application of Eq.(12), giving an output of 0.151, 0.113, and 0.103, for Imazalil, Thiabendazole, and Ortho-phenylphenol, respectively. The difference between the probabilities obtained from the randomly drawn sample (representative sample) and the population sample (ground truth) indicates that the probabilities fall within acceptable range for all pesticides. Here two points should be noted; the first is that if the validation method proposed here is performed independently more than once, the probabilities obtained will have high precision but it is unlikely to have the same value at every digit, for the fact that a random number generator is involved in the process, the second point is convergence of the probability obtained by the random drawn sample to the probability obtained by the ground truth is sufficient to imply the expected outcome (validity of the proposed method). The

difference between the probabilities obtained by the ground truth and the randomly drawn sample for each pesticide can be attributed to the exhaustive contemplation conducted by the probabilistic bootstrap sampling, opposed to direct computation using Eq.(12). The exhaustive contemplation, however, is an advantage rather than limitation in real case scenarios for the fact, CL is a continuous variable with a very low chance to encounter the same CL in any two randomly drawn samples (evidenced by the data from NVWA) with the ground truth generally unknown. Therefore, the probabilistic bootstrap sampling allows the contemplation of unobserved but likely case scenarios to be included in the computation.

## Validation Two

For the same reason as discussed in validation one section, a reference data that can be accepted as the ground truth, reflecting all possible CL to be encountered in a laboratory settings is compulsory. Therefore, for this particular validation step, all 4896 CL values obtained from NVWA dataset was extracted and assigned as the ground truth for a single hypothetical COI. This way, opposed to using a random number generator for such form of validation, a realistic value can be employed. In a similar fashion as in validation one, given the ground truth, a random draw was conducted that consisted of only 30% of the uniquely observed CL as representative CL values for the hypothetical COI (Fig.S4). That way, a worst case scenario can be contemplated and the robustness of the method can once again be tested.

The randomly drawn CL represents the large number of observations normally required to apply the proposed method. Given this random draw, the idea behind the validation step is, in a similar fashion as in the previous validation step, to access and prove if, in fact the $P(H_0)$ obtained by using the randomly drawn representative samples will converge to the $P(H_0)$ obtained by direct application of Eq.(5) to the ground truth (the entire NVWA data in this case).

Opposed to validation one, the ground truth is large in size, and obtaining the expected outcome would imply that the method proposed is independent of the population sample but only dependent on the representativeness of the unique CL values obtained (in this case only 30%). The analytical method boundaries simulated for demonstrative purposes and applied for validation one (Fig.S1) are also once again adopted for this validation method. Given the data and the method parameters, Fig.S5A-B depicts the classification (clustering) results and the posterior probability distribution obtained by applying probabilistic-bootstrap-sampling. The mean as the maximum likelihood estimate of the normally distributed posterior probability was calculated to be 0.172. Similarly, given the method parameters, $P(H_0)$ was also calculated for the ground truth by direct application of Eq.(5) which obtained a result of 0.204. The convergence of the probability obtained by using the representative random draw towards the probability of the ground truth implies the expected outcome and underlines the validity of the proposed method. It should be noted that increasing the percentage of the randomly drawn uniquely observed CL will obviously lead to more convergence (result now shown). Taking in to account the validation step was aimed at testing the approach at worst case scenario, the result obtained with just 30% random samples has been found to be sufficient enough to validate the method.

## Figure Captions

**Figure S1.** Graphical representation of simulated analytical methods performance boundaries.

**Figure S2.** Graphical representation of ground truth and random drawn samples. **Part A** Imazalil. **Part B** Thiabendazole. **Part C** Ortho-phenylphenol

**Figure S3. Part A-C** Graphical representation of classification (clustering) result of continuous concentration level (CCL), versus discrete concentration level (DCL) for example bootstrap samples for the purpose of validation one, for Imazalil, Thiabendazole, and Ortho-phenylphenol, respectively.

**Part D-F** Histogram representation of posterior probability distribution, obtained by exhaustive probabilistic-bootstrap-sampling for the purpose of validation one, for Imazalil, Thiabendazole, and Ortho-phenylphenol, respectively.

**Figure S4.** Graphical representation of ground truth and random drawn samples for hypothetical COI for the purpose of validation two.

**Figure S5. Part A** Graphical representation of classification (clustering) result of continuous concentration level (CCL), versus discrete concentration level (DCL) for example bootstrap samples for the purpose of validation two, for hypothetical COI.

**Part B** Histogram representation of posterior probability distribution, obtained by exhaustive probabilistic-bootstrap-sampling for the purpose of validation two, for hypothetical COI.

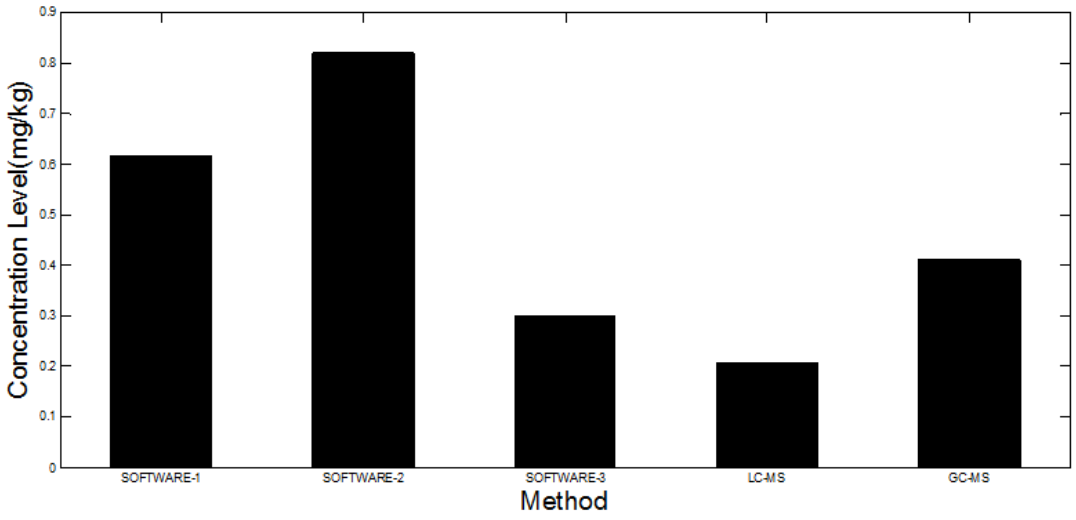


Figure S1

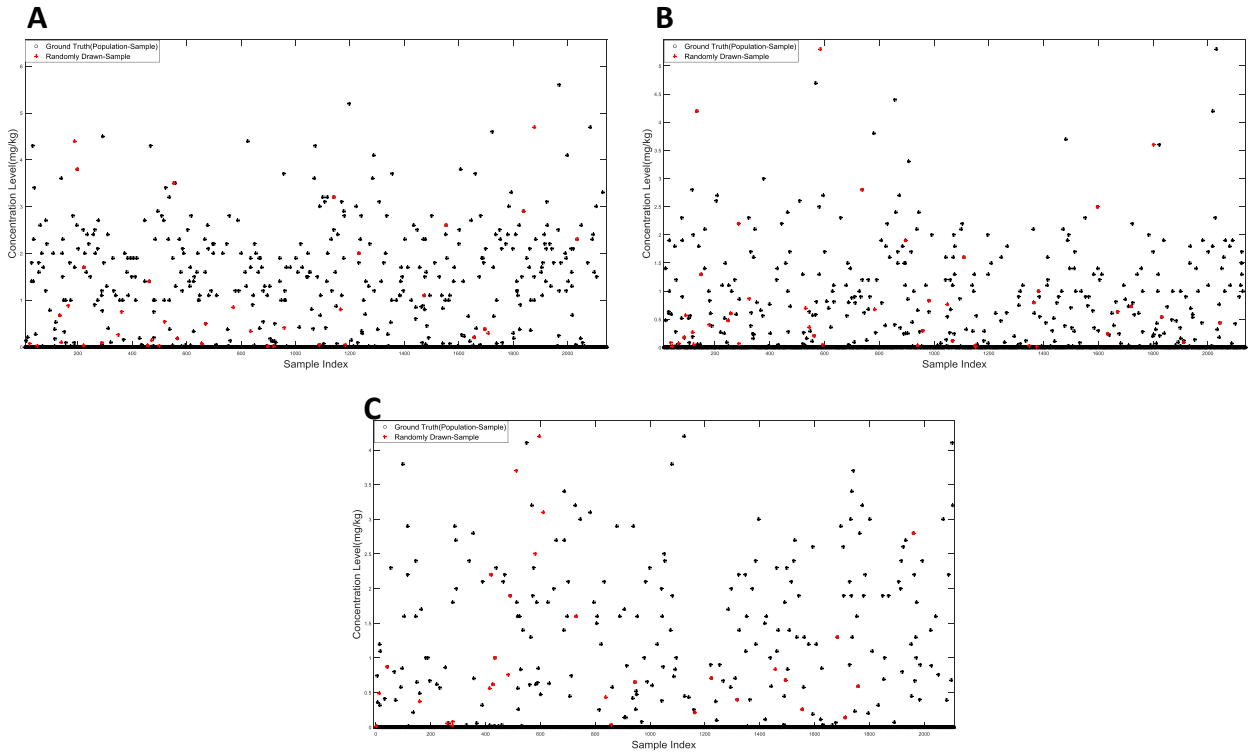


Figure S2

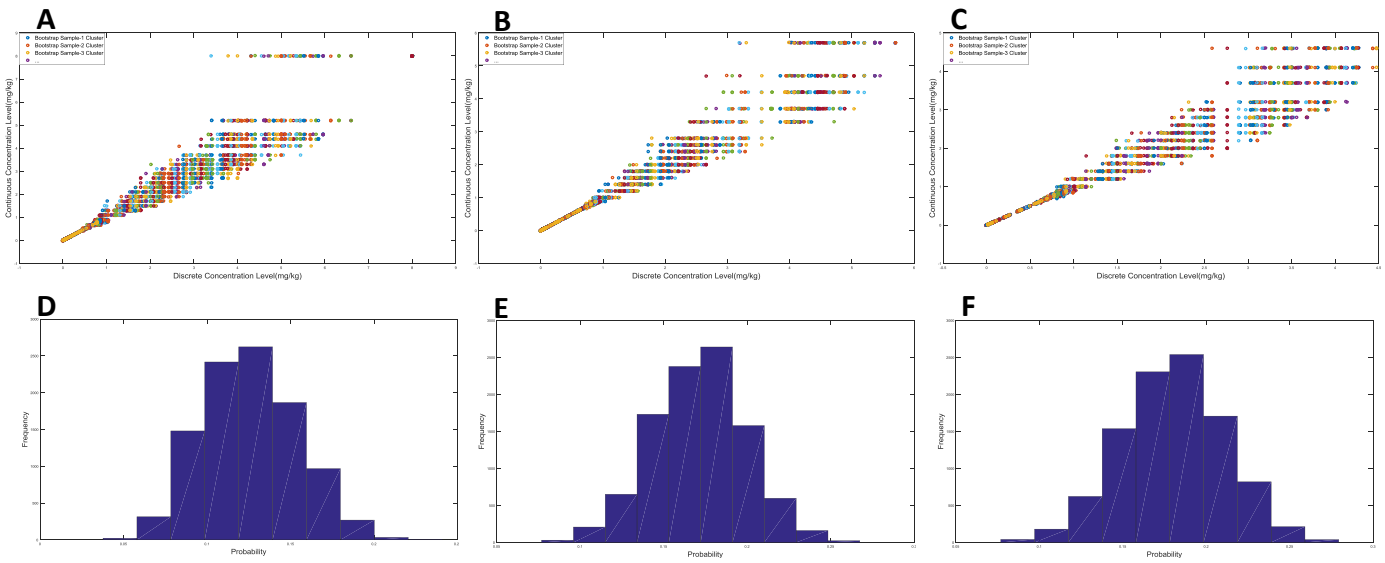


Figure S3

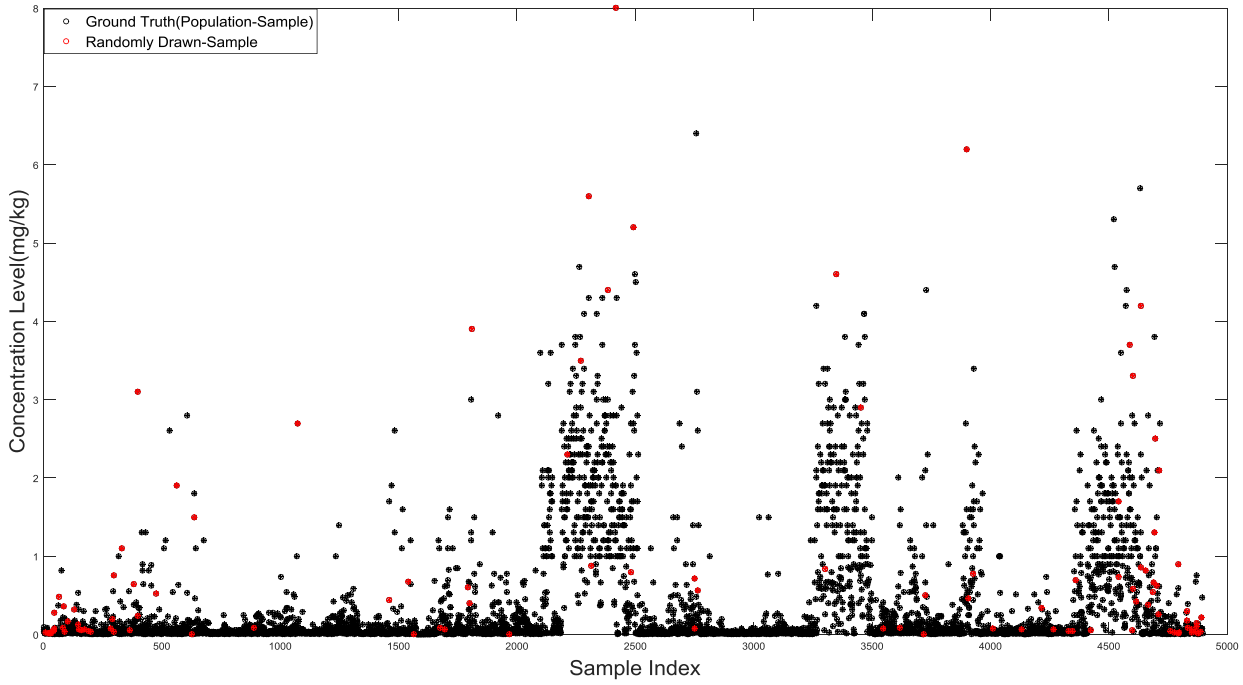


Figure S4

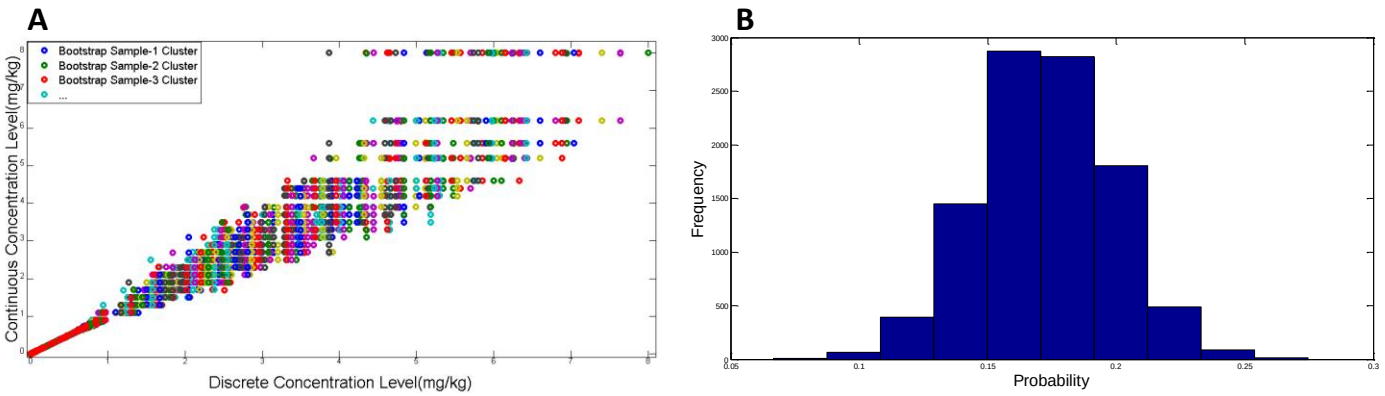
A
Bootstrap Sample-1 Cluster
Bootstrap Sample-2 Cluster
Bootstrap Sample-3 Cluster
...
Continuous Concentration Level(mg/kg)
Discrete Concentration Level(mg/kg)
B
Frequency
Probability

Figure S5