A STATISTICAL METHOD FOR THE ANALYSIS OF PCB PROFILES IN THE PRESENCE OF VALUES BELOW THE DETECTION LIMIT

W. Talloen[1], B. Vrijens[1], M. Vleminckx[2], A.De Cock[2], G. De Poorter[2], S. Srebrnik[1], L. Goeyens[1] and J. Willems[3]

[1] Institute of Public Health, Juliette Wytsmanstraat 14, 1050 Brussels, Belgium

[2] Belgian Federal Ministry of Agriculture, WTC3 Boulevard S. Bolivar 30, 1000 Brussels, Belgium

[3] Heymans Institute of Pharmacology, Ghent University, Medical School, 9000 Ghent, Belgium

Introduction

On May 28, 1999, the fact that PCBs and dioxins had entered the Belgian food supply through contaminated animal feed was made public. In February and March, Belgian farmers had already witnessed a serious drop in egg production, poor egg hatching and increased mortality of chicks. From June 1999 on, an intensive monitoring program was launched in order to evaluate the presence of PCBs/dioxins in food items. Since the origin of the dioxins appeared to be a contamination with PCBs—most probably Aroclor 1254 and 1260—screening of some 50000 samples started with an analysis of the marker PCBs. We thought it of interest to perform an analysis of the PCB profiles found in order to obtain a better insight in the original contamination(s).

Residue research often faces analytical results that fall below the assay's limit of quantification. These so-called "below detection limits" (BDLs) present a serious interpretation problem[1] because the left tail of the PCB distribution pattern is censored. Leaving the censored data out, as well as substituting them by one single value, e.g., one-half of the reported limit, will bias the estimation of the distribution parameters[1]. Hughes[2] has shown on the basis of simulations that these procedures can produce estimates with significant bias, especially as the proportion of data that is censored increases. We here propose some statistical methods in order to handle censored data in an appropriate way.

Data

The monitoring program resulted in a data set of PCB profiles in animal feed, eggs, chicken, pork, beef, milk and derived products. Standard PCB analyses consisted in the determination and, if possible, quantification of 7 PCB markers (PCBs 28, 52, 101, 118, 138, 153 and 180). Hence, each analysis i combines information on sample characteristics such as sampling moment, location, laboratory, ...), with a vector $Y_i = (y_{ij}, j=1,...,7)$ as a function of the presence of each of the 7 marker PCBs.

In order to illustrate our approach we simulated a data set of 20 samples, generated according to a multinomial distribution by letting a random total amount (n_i) to be distributed among the 7 possible congeners according to a probability vector Π. Furthermore, we generated an extra-multinomial variation by allowing the probability vector Π_i to vary across samples as illustrated in table1.

Table 1: Example of 2 profiles (sample *i* and *i'*) as an illustration of the simulated data structure

PCB		28	52	101	118	138	153	180	
pop. prob.(Π)		0.05	0.05	0.2	0.1	0.2	0.25	0.15	
indiv. prob.(Π_i)	*i*	0.04	0.058	0.196	0.106	0.178	0.262	0.16	
	i'	0.034	0.04	0.22	0.076	0.208	0.258	0.164	
conc.	*i*	3*	9*	43	25	45	82	45	$n_i = 252$
	i'	9*	14	61	22	46	66	46	$n_{i'} = 264$
conc. with detection limits	*i*	<13	<13	43	25	45	82	45	
	i'	<13	14	61	22	46	66	46	

Methods

The sum of the 7 PCB congeners can be considered as response variable and a log normal distribution can be assumed. A regression versus potential covariates can then be carried out, acknowledging for below detection limit values via a likelihood term by integrating over the possible censored range of observations.

An alternative method which is implemented in standard statistical packages, considers the inverse of the concentrations in a survival analysis framework. When investigating the vector of PCB responses $Y_i = (y_{ij}, j=1,\ldots,7)$, a multivariate extension of the above described method could be implemented.

To study the possible sources and number of sources of food contamination, we will mainly investigate how the total observed amount of PCB (n_i) is distributed among the 7 congeners in each sample *i*. In other words we want to make inferences around the distribution of the probabilities (Π_i), instead of the concentrations themselves. Intuitively one can compute for each sample the percentages and make use of the central limit theorem to apply traditional multivariate techniques like principal component analysis, cluster analysis, discriminant analysis and biplots.

We face, however, the problem of BDL, which will typically bias the estimation of the percentages. Neglecting samples with BDL values or setting them equal to zero, to the detection limit or the half of it, will bias the profiles. As an illustration we introduced a detection limit of 40 in our simulated data set of 20 samples, so that 50% of the data became below the detection limit. Figure 1 shows the profiles of the simulated data where we set BDL values equal to zero, equal to the detection limit or equal to half the detection limit.

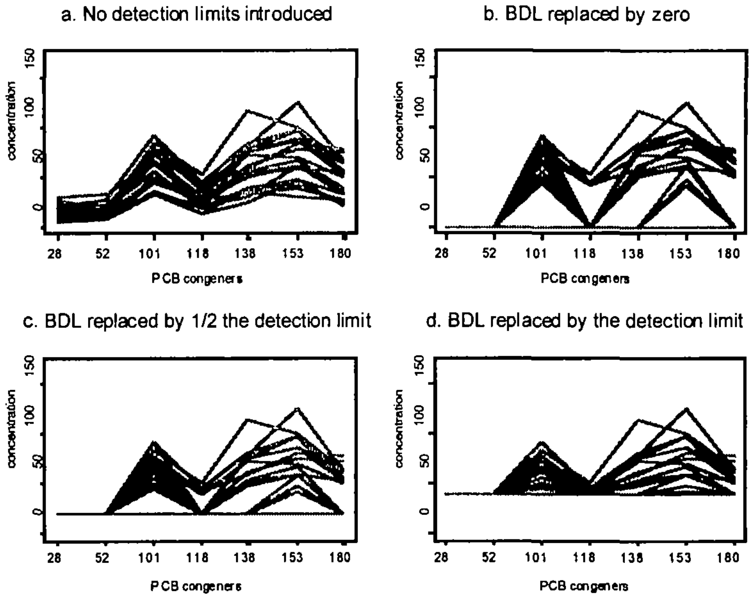

Figure 1: Profiles of 7 PCB markers using a simulated data set (sample size=20). Figure a shows the original data without the presence of detection limits. Figure b, c and d are profiles of the same data set with the introduction of detection limits equal to 40. The BDL values are replaced by zero in figure b, by one half the detection limit in figure c and by the detection limit in figure d.

Although the samples are all issued from the same probability vector, a quite different pattern arises after BDL values are introduced, each approach giving a different result. Furthermore, the introduction of BDL values considerably decreases the variability of the observations.

We suggest to model the outcome $Y_i = (y_{ij}, j=1,...,7)$ as if it was issued from a multinomial distribution. In order to acknowledge for extra-multinomial variation that can be expected to arise from concentration data, multivariate random effects were added. This was implemented in a MCMC framework by means of the Multinomial-Poisson transformation (MP), where censored observations are included in an appropriate way by introducing an additional parameter for each censored observation. In this way, all conditional distributions are unchanged from the analysis with no censored observations and a new set of conditional distributions for the set of censored observations is introduced[3]. An additional advantage of this approach is that it provides estimates of the full posterior distributions of the parameters. This MCMC approach is carried out in a Bayesian inference framework using Gibbs sampling as simulation technique in the Software WinBUGS 1.2 [4].

Results and discussion

We estimated the percentages of our simulated data set by means of two statistical methods, the replacement of BDL values by zero and the MCMC approach.

Table 2: **Estimates of mean and standard deviation for the same data set using two different statistical models.**

PCB	used probabilities	Substituting BDL by zero			MCMC approach		
		mean	±	standard deviation	mean	±	standard deviation
28	0.05	0.00678		0.002124	0.04273		0.003556
52	0.05	0.0201		0.007724	0.0482		0.00374
101	0.2	0.2092		0.007643	0.1962		0.006988
118	0.1	0.1155		0.005013	0.1071		0.005856
138	0.2	0.2226		0.006964	0.2066		0.007657
153	0.25	0.268		0.009176	0.2518		0.008586
180	0.15	0.1578		0.005687	0.1472		0.006439

It appears that the percentages of PCB 28 and PCB 52, which are the PCB congeners presenting most BDL values, are underestimated when substituting the BDL by zero. The MCMC approach consistently estimates all PCB congeners correctly. This stresses the importance of implementing BDLs in a correct way. Going back to the original data base of 50000 chemical analyses, we face the problem that the data set contains values censored at more than one detection/reporting limit. The reasons are multiple: the quantification limits were not the same for all the congeners, different labs performed the analyses and within a same lab the detection limits became lower over time. Furthermore, there are many covariates involved in this data set: type of matrix, time of collection, laboratory identity and others. These two difficulties are implemented in our methodological approach. An important difficulty, however, remains. The implementation via Gibbs sampling is very computer intensive. Each censored Y_{ij} is iteratively sampled for the full multivariate distribution until it gets a value below the censoring point. The model requires a lot of computer time when the censored range is restricted. Implementation on a UNIX platform and model reparametrisation could eventually result in significant improvement but has not yet been tested.

References

1. Helsel D.R. (1990) Less than obvious: Statistical treatment of data below detection limit; Environ. Sci. Technol.; 24 (12):1766-1774.
2. Hughes J.P. (1999) Mixed effects models with censored data with application to HIV RNA levels; Biometrics; 55:625-629
3. Baker S.G. (1994) The Multinomial-Poisson transformation; The Statistician; 43 (4):495-504.
4. Spiegelhalter D.J., Thomas A., Best N.G. and Gilks W.R. (1995) BUGS: Bayesian Inference Using Gibbs Sampling, version 0.50. MRC Biostatistics Unit, Cambridge, UK.