Treatment of Values Below the Detection Limit in Correlation Analysis of Chlorinated Dioxins and Related Compounds

Kaune, A.
GSF-Forschungszentrum für Umwelt und Gesundheit GmbH, Projektgruppe Umweltgefährdungspotentiale von Chemikalien, Ingolstädter Landstraße 1, 8042 Neuherberg, Germany

Abstract

Concentrations of polychlorinated dibenzo-*p*-dioxins and dibenzofurans (PCDD/F) in incinerator exhaust gas were correlated with concentrations of potential indicator parameters which in some measurements were below the limit of detection (LOD). Results indicating that a correlation was significant when treating values below LOD as missing were confirmed for some data sets, but had to be modified for others when the correlation analysis was repeated after the values below LOD were replaced by data between LOD and zero. I therefore propose to treat values below LOD as missing in the final regression equation and to check whether the significance of the correlation also holds for the entire data set including non detected concentrations. This can be done by replacing values below LOD by a value smaller than the minimum of detected concentrations and by testing the significance of the rank correlation coefficient.

Introduction

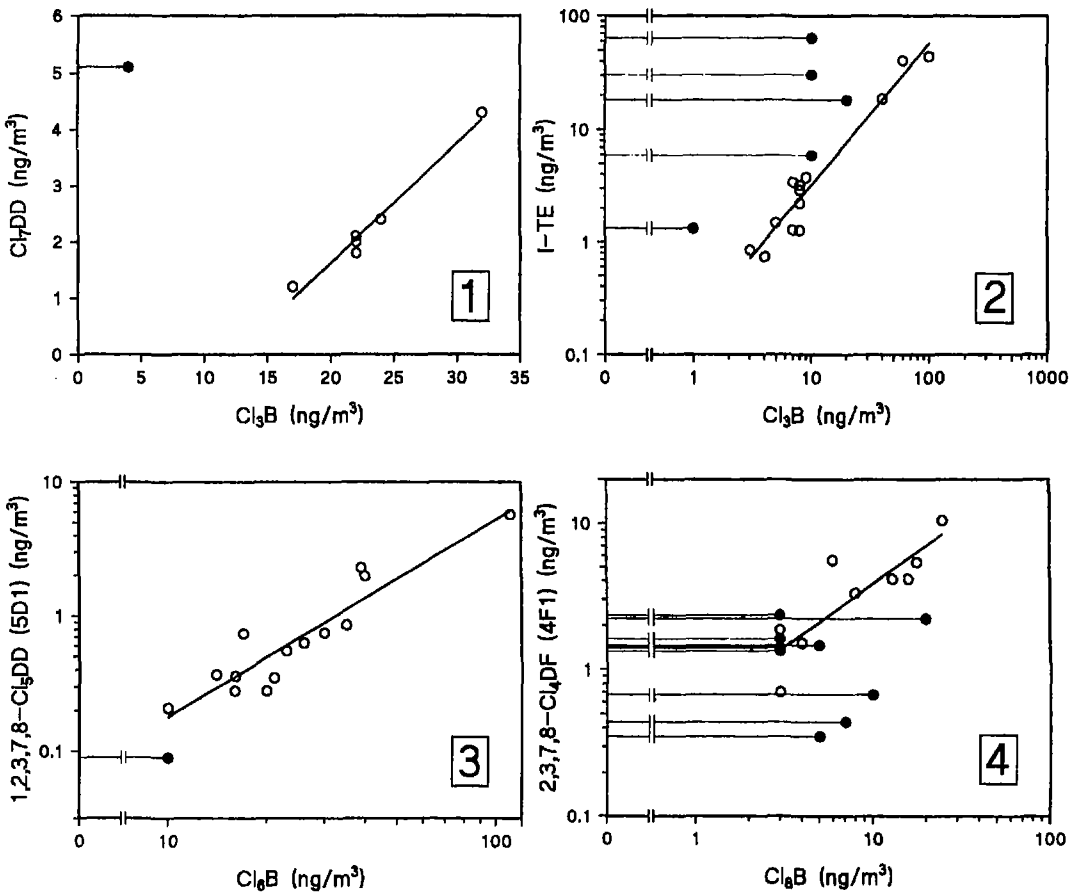
There is a still increasing interest to find indicator parameters for PCDD/F concentrations and toxicity equivalents in incinerator exhaust gas. Multiple regression and partial least squares regression were used to establish the relationship between PCDD/F and indicator parameters[1]. Linear regression equations have also been reported[2]. The concentrations used as variables in the statistical analysis may sometimes be below LOD and were then treated as missing values[1]. However, the effect of this treatment on the significance of the correlation and regression has not yet been examined. This will be done in the present paper.

Materials and Methods

Data measured in different parts of the flue gas cleaning system and in the stack gas of a hazardous waste incinerator were taken from a work which I have reported on in reference[3]. Calculation of the Pearson correlation coefficient (R) and of the parameters of a regression equation requires the variables to be normally distributed. This was usually obtained by taking the logarithm to base 10 of the values and was tested employing the normal probability plot correlation method. If the normality assumption did not hold on the 0.01 level of significance Spearman's rank correlation coefficient (r) was computed. All Computations were done using SAS procedures[4].

RESULTS AND DISCUSSION

Example 1. The correlation of heptachlorodibenzo-*p*-dioxin (Cl_7DD) vs trichlorobiphenyl (Cl_3B) (Figure 1) was highly significant (R = 0.988, P = 0.0002). One measurement was excluded in which Cl_3B was not detected and the concentration of Cl_7DD was 5.1 ng/m^3. Given this Cl_7DD level, Cl_3B should clearly be detectable if the regression equation depicted in Figure 1 was valid for the entire data set. As a consequence, the significance of the correlation of Cl_7DD vs Cl_3B seems to be artificial, purely resulting from treating the value below LOD as missing. A statistical confirmation of this conclusion can be obtained by repeating the correlation analysis after the value below LOD was replaced by any positive value equal to or lower than LOD. Using e. g. LOD or LOD/2 yielded $R = -0.277$ and $R = -0.318$, respectively. Both correlation coefficients were insignificant on the 0.05 level of significance. The same was true for the rank correlation coefficient (r = 0.185, P = 0.691) using any positive value ≤ LOD. It may be argued that in this example the analysis of Cl_3B was erroneous. However, example 2 demonstrates that the same problem as in example 1 can occur with a larger number of non detected concentrations all of which can hardly be assumed to be erroneous. Cl_3B was not approximated very well by the normal distribution (P = 0.011) but the correlation coefficient (R = 0.969) was highly significant ($P < 0.0001$). The correlation did not include measurements with non detectable Cl_3B concentrations for which the corresponding NATO/CCMS toxicity equivalents (I-TE) ranged from 1.32 to 63.83 ng/m^3. For these values, Cl_3B should have been detected, even if considering that LOD was relatively high for some measurements, on the assumption that I-TE significantly regressed on Cl_3B. A statistical confirmation of rejecting this assumption cannot be as easily obtained as in example 1 because the correlation was still significant when LOD (R = 0.761, P = 0.0002) or LOD/2 (R = 0.608, P = 0.0058) was presumed for non detected Cl_3B concentrations. However, values below LOD were replaced by LOD × f where f is a factor ranging from 1 to 0. Using $f = 0$ is equivalent to treating the value as missing since data were log transformed and the logarithm of 0 is not defined. Decreasing f from 1 to 0.05 and thus decreasing the value by which non detected concentrations were substituted decreased the correlation coefficient (Figure 5). This decrease was regarded as a hint that the regression of I-TE on Cl_3B, although significant for detected values, was insignificant for the parent population. Unlike examples 1 and 2, examples 3 and 4 show that the results of a correlation analysis were confirmed even if there were, respectively, one or many values below LOD. Example 3. The correlation of 1,2,3,7,8-pentachlorodibenzo-*p*-dioxin (5D1) vs hexachlorobiphenyl (Cl_6B) (Figure 3) was highly significant ($P < 0.0001$), either excluding one value below LOD or replacing it by LOD or LOD/2 (R = 0.931, and R = 0.943, respectively). Furthermore, the rank correlation coefficient (r = 0.892) was highly significant for all values < LOD. Example 4. Replacing values below LOD by LOD × f with f decreasing from 1 to 0.2 increased the correlation coefficient (Figure 5). This is opposite to example 2 and was taken as a hint that 4F1 in fact depended on Cl_8B.



Figures 1–4. Illustration of examples 1–4. Values < LOD are represented by closed circles (indicating LOD) with a bar to the left (indicating the range of possible values).

CONCLUSIONS

The concentration level of a substance which was not detected may be just a little lower than or it may be some orders of magnitude lower than LOD. Due to this lack of quantification, values below LOD should not be included in correlation and regression analyses, i. e. be treated as missing values in the *final* regression model. However, values below LOD contain more information than missing data (which do not contain any information at all). This information should be used to verify whether the correlation and regression derived for detected values is also valid for the entire data set including the values below LOD. This can be done by plotting all data and by visually examining whether values below LOD fit to the regression line of the detected values. On a statistical basis, correlation and regression analysis should be repeated after replacing the values below LOD by a few values ranging from LOD to zero (or close to zero in case of log transformation). Then, the distribution of the data may not be normal any more. In this case, rather than calculating the Pearson correlation coefficient the non-parametric Spearman rank correlation coefficient has to be determined. The most promising and least laborous way to obtain reliable results for all the examples presented is to use any value smaller than the minimum of detected data. The ranks of data and thus the rank correlation coefficient (*r*) will then be constant. If *r* is insignificant there is most likely no relationship between both variables and no dependence of *y* on *x*. In the examples shown, the independent variable was examined. It can, however, be assumed that the same problem also arises and that the proposed solution also applies for a dependent variable containing values below LOD and for regression techniques other than the linear one with one independent variable used in the examples above.

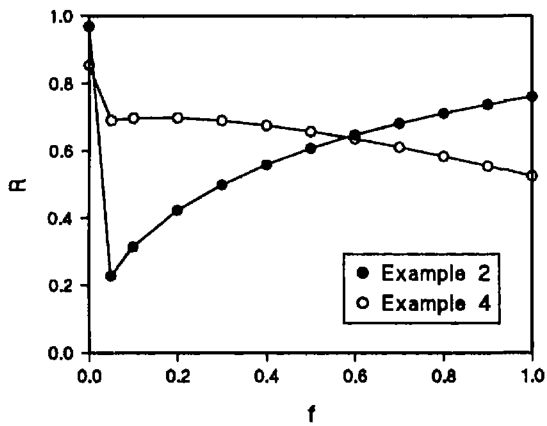


Figure 5. Pearson correlation coefficients (*R*) for examples 2 and 4 in which values < LOD were replaced by LOD × *f*. In example 2, correlation coefficients < 0.42 were insignifcant on the 0.05 level of significance.

References

1 Öberg T, Bergström J. Emission and chlorination pattern of PCDD/F predicted from indicator parameters. *Chemosphere* 1987; 16: 1221–1230.

2 Oehme M, Manø S, Mikalsen A. Formation and presence of polyhalogenated and polycyclic compounds in the emissions of small and large scale municipal waste incinerators. *Chemosphere* 1987; 16: 143–153.

3 Kaune A, Lenoir D, Nikolai U. Estimating toxicity equivalents in the stack gas of a hazardous waste incinerator from the concentrations of chlorinated benzenes and biphenyls. Submitted to Dioxin '93, 13th International Conference on Chlorinated Dioxins and Related Compounds, Vienna, 20–24 September 1993.

4 SAS Institute Inc. *SAS Procedures Guide*, Version 6, Third Edition, Cary, NC: SAS Institute Inc., 1990.