



2-FUN

*Full-chain and UNcertainty Approaches for Assessing Health Risks in
FUture ENvironmental Scenarios*

**FP6 Project-2005-Global-4
Integrated Project - Contract n°: 036976**

– REVIEW AND TESTS OF METHODS FOR ROBUST DATA TREATMENT –

Due date of delivery: *29/02/2008*

Actual submission date: *04/03/2008*

Start date of the project: *01/02/2007*

Duration: *48 Months*

Lead contractor organisation name for this deliverable: *IPH*

Project co-funded by the European Commission within the Sixth Framework Programme (2002-2006)	
Dissemination Level	
PP	Restricted to other programme participants (including the Commission Services)

PROPRIETARY RIGHTS STATEMENT

This document contains information, which is proprietary to the 2-FUN Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to any third party, in whole or in parts, except with prior written consent of the 2-FUN consortium.



Document Information

Document Name REVIEW AND TESTS OF METHODS FOR ROBUST DATA TREATMENT
ID D1.4 robust data treatment report v6.doc
Revision Version 6
Revision Date 05/03/2008
Author P. KOVANIC / IPH & P. CIFFROY / EDF

Approvals

	Name	Company	Date	Visa
Author	P. KOVANIC	IPH	04/03/2008	P. Kovanic
Co-Author	P. CIFFROY	EDF	04/03/2008	P. Ciffroy
WP Leader	A. MARCOMINI	UNIVE	06/03/2008	po E. Giubilato
Coordinator	F. BOIS	INERIS	06/03/2008	F. Bois

Documents history

Revision	Date	Modification	Author
Version 0	18/01/2008	Template made available to the WP leaders	F. BOIS
Version 1	27/01/2008	First version	P. KOVANIC
Version 2	07/02/2008	Revised version upon WP1 leader comments	P. KOVANIC
Version 3	14/02/2008	Revised version upon WP2 leader comments	P. KOVANIC
Version 4	21/02/2008	Final draft version	P. KOVANIC
Version 5	03/03/2008	Deliverable made available to the coordinator	P. KOVANIC
Version 6	05/03/2008	Final release, in the standard format	F. BOIS



Contents

INTRODUCTION.....	4
BRIEF DESCRIPTION OF TESTED METHODS	4
1. REQUIREMENTS TO DATA TREATMENT METHODS FOR HRA	4
TESTS ON TYPICAL CASES MET IN RISK ASSESSMENTS.....	12
1. LIMITED SET OF HOMOGENEOUS DATA	12
2. DATASET CONTAINING SEMI-QUANTITATIVE DATA (LOWER THAN THE DETECTION LIMIT).....	15
3. DATASET CONTAINING OUTLIERS OR UP-CENSORED DATA	16
4. TESTS OF REGRESSION METHODS.....	17
5. TESTS OF CORRELATION COEFFICIENTS	18
DISCUSSION	19
1. PROBABILITY AND DENSITY DISTRIBUTIONS.....	19
2. ESTIMATION OF DISTRIBUTION FUNCTIONS OF INCOMPLETELY DEFINED (CENSORED) DATA	20
3. ROBUST REGRESSION MODELS.....	20
4. CORRELATION COEFFICIENTS	22
CONCLUSIONS	22



INTRODUCTION

Exposure and effects models used in risk assessment are generally characterised by high uncertainties resulting from: (i) uncertainty of model parameters and (ii) lack or uncertainty of input data.

Exposure and effects models require indeed many parameters that describe the physical, chemical or biological behaviour of stressors in the environment (e.g. Transfer Factors to biological compartments like vegetables, milk, animals... Degradation and Diffusion rates...), their potential transformations during human procedures (e.g. Elimination rates during cooking...) and their behaviour in the human body (e.g. Toxicokinetic rates). For a given exposure model as defined by its structure (compartments and processes), a major source of the output uncertainty arises from the uncertainty of these key parameters. To address this problem, several methods like Weighted Bootstrap approach, Bayesian approaches and QSAR are developed in the 2-FUN's WP2 (see 2-FUN's Deliverable 2.3).

Exposure models use also input data regarding the emissions of pollutants from anthropogenic activities (e.g. from industrial plants) and/or concentrations of hazardous substances in the environment (measured or estimated in water, air, soil or biological compartments). However, the accurate treatment of such data before their introduction in exposure models is difficult. Indeed, for many hazardous substances that are released to or present in the environment at trace or ultra-trace levels, their analytical determination is cost- and resource-intensive and thus, for many environmental stressors, the number of useable data produced by monitoring programs remains low. To date, such uncertainties have generally been ignored in exposure assessment, as commonly data on emissions and/or concentrations in the environment are (i) either averaged (from a very limited set of data), or (ii) censored or misinterpreted. The latter is often the case for values below the analytical detection limits, or outliers (i.e. values that cause surprise in relation of the majority of the dataset). However, the occurrence of temporal and/or spatial peaks can be of high concern for health risks of individuals. Therefore, on one part, when these data are ignored, exposure assessments can lead to under- or over-predictions, but on the other part, a single outlier can completely upset the mean and variance of the dataset and lead to misinterpretations.

The term "statistical methods" is generally understood as "methods of classical statistics" that are strongly sensitive (in some cases non-robust) with respect to strong data uncertainties and/or outliers frequently met in environmental datasets. More robust methodologies are then needed to improve the 'upstream' data treatment before their introduction in exposure models, in particular to:

- robustly interpret limited sets of data (typically a limited set of measurements of a given pollutant in an environmental media);
- include in the data treatment semi-quantitative data such as values below the detection limits;
- incorporate in the data treatment values which are commonly censored (e.g. outliers);
- fill data gaps from correlations with environmental factors (typically correlation between a well-monitored pollutant and a bad-monitored pollutant respectively).

Some methods that potentially provide rationale and theory-based information for each of these concrete questions were tested, compared and discussed in the present report. In order to give a wide overview of potential approaches applicable to small samples of strongly dispersed data, both robust statistical methods and an alternative mathematical approach were tested and illustrated on different sets of data.

BRIEF DESCRIPTION OF TESTED METHODS

1. Requirements to data treatment methods for HRA

To select some methods able to satisfy needs commonly met in environmental and health risk assessments, some pragmatic requirements were previously defined:

- 1) The methods must be available in the computerized form, as a software package;



- 2) The methods must serve to both general and special needs of the HRA:
 - a) to solve both one-dimensional (marginal) problems and multi-dimensional (MD-) problems;
 - b) to be efficient in application to small samples of data;
 - c) to be objective, not relying on subjective a priori assumptions on features of the data and/or of observed processes;
 - d) to not rely on special requirements to ways of collecting and measuring data like randomness, independence, stationarity or homogeneity of data;
 - e) to be applicable to making use of information borne by incompletely defined data (low- or up-censored data, interval data);
 - f) to respect finiteness of the real world by efficient estimating and using the finite bounds of data domains;
 - g) to characterize the data samples not only by their location and scale, but also more completely by their probability and density distribution functions.
- 3) The methods must be robust with respect to “bad” data, which can be of different nature:
 - Outliers (extreme data or data from the non-homogeneous subsamples),
 - Inliers (inner disturbances or noises of data samples).
- 4) Distinguishing between “good” and “bad” data followed by adequate discrimination and evaluation of the data quality is a desirable feature of the data treatment method.

1.1. Selection of methods to be tested

To test computing methods, one needs usable functions running on a computer (according to the first criteria previously defined in 1.1). However, the way from an idea of a method to such a computerized function is long. An algorithm is to be created and converted into a program. Successful application of the program can depend on setting of some parameters. Moreover, some problems of stability and/or reliability of the functions can appear. The setting and operation problems should be overcome by the method's author or at least under his supervision. When a user U wants to compare methods of two authors A1 and A2 without having their authorized programs, he must try to create his (U's) program realization of the methods. In the case of some negative results of his test, he will face a risk of being subjected to (may be) ungrounded accusation for biasedness or for a not perfect realization of the A1's or A2's idea. This is a good reason to compare only authorized methods presented as programs of functions. It is worth mentioning how poor is the choice of such functions applicable within the famous commercial computing systems in spite of the abundance of the theoretical literature.

It remains thus to rely on the “invisible hand of the market” that the methods accepted by the commercially available data analysis systems are the most recommendable. This is the reason why methods already available under S-PLUS¹ will be applied below to test and compare several methods.

1.2. Statistical kernel methods available under S-PLUS

The standard probability distributions assumed in statistics need some parameters to be completely defined. Their applications are sometimes called “parametric distributions”. This is not quite precise because all non-standard methods of estimating the probability and density distributions need some parameters, at least the scale parameter, to be applicable.

Methods getting distributions directly from data are then called “nonparametric”. The needs of practice motivated development of such methods among which the “kernel estimates” play an important role.

The idea to apply additively composed kernels (having a prescribed form and located above the individual data locations) to estimation of probability density functions was firstly published by E.Parzen². Although new in statistics, such an idea was previously known in theoretical physics and theory of differential equations as the

¹ S-PLUS is a registered trademark of Insightful Corp., Seattle, Washington, USA

² Parzen E.: On estimation of a probability density function and mode, Ann. Math. Statistics 35 (1962), 1065-1076.



method of Green's functions and in linear systems theory as Duhamel's or convolution integrals. However, there is an essential difference between the Parzen's and the previous methods related to the form of kernels. In application to physics, e.g. to the problems of gas diffusion, the kernels' form was uniquely determined by the physical parameters of the gas. In application of Duhamel's integral to e.g. vibration of a string, the kernel's form resulted from mechanical features of the string. Unlike this, Parzen's kernels had a relatively arbitrary form constrained only by the requirement of convergence to normal distribution in the case of increasing the number of data. Parzen has established such conditions and proved the convergence under these conditions. The principles of the kernel approaches are the following:

If $x_1, x_2, \dots, x_N \sim f$ is an independent and identically distributed (i.i.d.) sample of a random variable, then the kernel density approximation of its probability density function is:

$$f_h(x) = \frac{1}{Ns} \sum_{i=1}^N K\left(\frac{x - x_i}{s}\right)$$

where K is a kernel function and s is the bandwidth (smoothing parameter). The determination of a kernel density function needs thus the definition of the kernel function and of the bandwidth s . The kernel function must result in a probability density function, and thus its integral is one. Besides, the kernel function ensures that the average of the corresponding distribution approaches (in limit case) that of the sample used.

In S-PLUS, there are four forms of kernel functions³:

- Gaussian: $K(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$ (the kernel used in Fig.1).
- Cosine: $K(x) = \frac{\pi}{4} \cos\left(\frac{\pi}{2} \cdot x\right)$ if $|x| \leq 1$, else 0.
- Rectangular: $K(x) = \frac{1}{2}$ if $|x| \leq 1$, else 0.
- Triangular: $K(x) = |1 - x|$ if $|x| \leq 1$, else 0.

Data of a sample can have different spread, characterized by the scale parameter S . Variable x in the formulae above can be thus written as $X \times S$, where X is the data item's value. Scale parameter thus determines the kernel's *width* (the bandwidth of the smoothing).

The form of kernel density can lead to functions presenting a variable number of modes according to the kernel function and *width* selected for the analysis. The only kernel defined and differentiable over the entire interval from 0 through infinity is the Gaussian. Its application to the sample of seven real data is shown in Fig.1. for different values of the parameter *width*.

Application of all the kernels available in S-PLUS to the same data sample with the default width is demonstrated in Fig.2.

A certain problem is that the number of kernels' forms satisfying the Parzen's conditions is infinite and there is no recommendation related to the optimum form. A user has to choose the kernel's form on his responsibility, without a hint of an optimization principle.

In S-PLUS, the user has also to choose not only the form of the kernel but also its width. There are following recommendations given with respect to the parameter **width** of the window:

This may be a numeric value, a function to apply to x , which returns a bandwidth, or a character string specifying, which built-in bandwidth method to use. Available bandwidth methods are histogram bin (hb), normal reference density (nrd), biased cross-validation (bcv), unbiased cross-validation (ucv), and Sheather & Jones pilot estimation of derivatives (sj). The standard error of a Gaussian window is width/4. For the other windows, width is the width of the interval on which the window is non-zero.

³ Venables, W.N. and Ripley, B.D. (1997) *Modern Applied Statistics with S-PLUS*, Second Edition. Springer-Verlag.

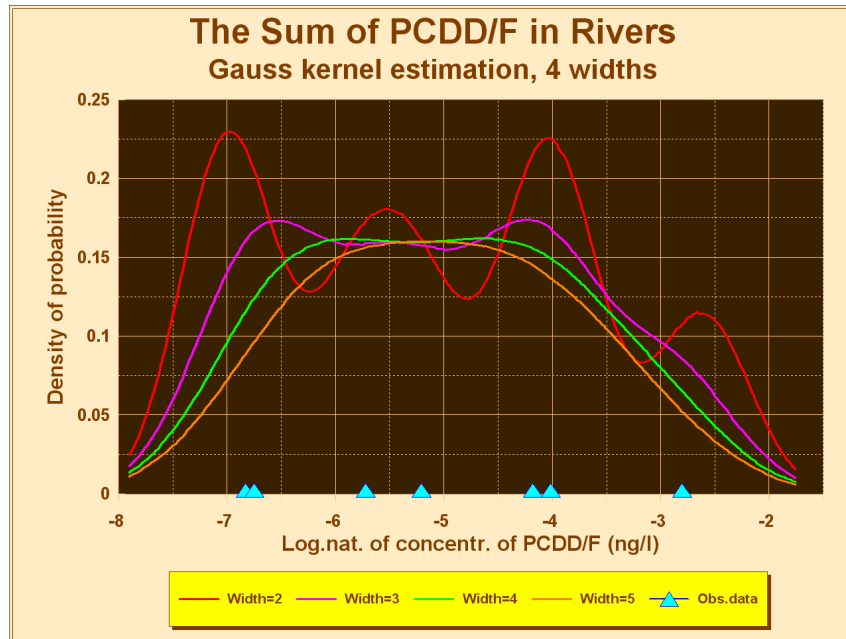


Figure 1: Statistical estimates of density function using the Gauss kernel and different widths

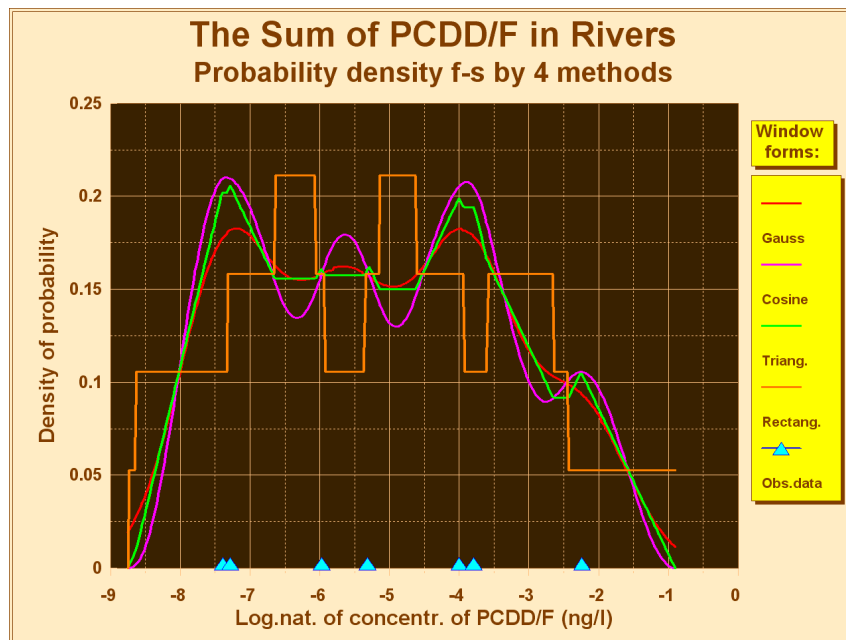


Figure 2: Kernel estimates of density function for logarithmic data

1.3. Robust estimation of the regression models

To fill data gaps from relationships with environmental factors, it is classical in HRA to use regression models. However, in the presence of outliers (as in many data samples used in HRA), 'classical' least squares estimation can be biased and are not robust. As in many cases it is precisely the outliers that are of interest (e.g. peaks in the



emissions or concentrations of pollutants in the environment), robust methods, using alternative estimates of the error of the regression, are needed.

S-PLUS system makes robust statistical models available of the M-type by means of the function *rreg*, which calls several auxiliary functions. The name of the method reminds the powerful statistical method of maximum likelihood, which can be here generalized to other ‘maximum-likelihood type’ estimators.

The routine *rreg* uses iterative Reweighed Least Squares method to approximate the robust fit, with residuals from the current fit passed through a weighting function to give weights for the next iteration. There are several possible weighting functions, and the user is free to create own function. The *weight* functions (frequently called ***influence*** functions) that are available are: *wt.andrews*, *wt.bisquare*, *wt.cauchy*, *wt.fair*, *wt.hampel*, *wt.huber*, *wt.logistic*, *wt.median*, *wt.talworth*, *wt.welsch*. Some of the names of weighting functions remind personalities of their famous authors, who not only published own methods but also recommended values of some function’s parameters.

The ***weight*** is a positive number usually less or equal to one dependent on the error of an equation of the equation system in the current iteration. The weight is given to the equation to decrease its impact on the result. The forms of the weighting functions are based on some statistical assumptions related to the data features. All weighting functions have the form of $w = f(\text{resid})$ where *resid* is a vector of residues/equation errors normalized by the scale parameter $\text{scale} = \text{median}(\text{abs}(\text{resid}))/0.6745$.

There are only five continuous and differentiable weight functions among ten mentioned statistical ones. Five non-differentiable functions are shown in Fig.3. The classical (OLS, Ordinary Least Squares) method can be considered as a special case applying fixed weights all equaling one.

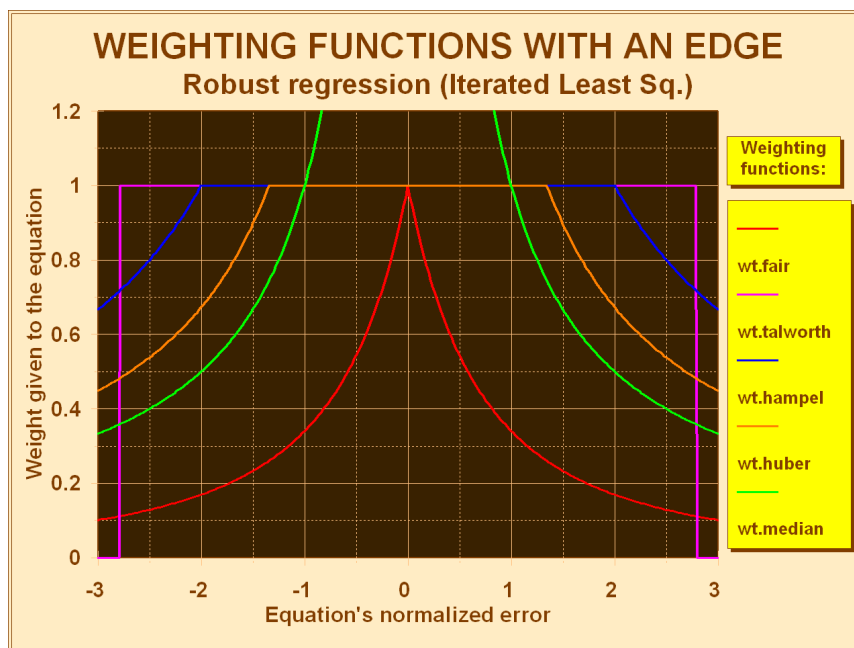


Figure 3: Five non-differentiable influence functions of the robust statistics

Three of these (*wt.talworth*, *wt.hampel* and *wt.huber*) apply the full weight 1 to equations, the error of which is within a fixed interval. This introduces an artificial a priori discrimination of data into “perfect” and “imperfect”.

Method *wt.median* compensates the influence of the worse data by giving the weight of $1/|\text{error}|$ to the data having a small absolute error.

Method *wt.fair* also has a singularity for a zero error. However, two points are worth mentioning:

- a) The rate of decreasing the weight of this method exceeds that of all others statistical methods available in S-PLUS;
- b) Just this method gave the best results of the tested statistical methods.

Effects of the continuous influence functions of robust statistics are shown in Fig.4.



Functions *wt.andrews* and *wt.bisquare* coincide; they are marked by the red line and vertical marks in Fig.4. All the demonstrated continuous functions have a finite curvature for all data errors and differ only by the local forms and especially by the drop of the weight.

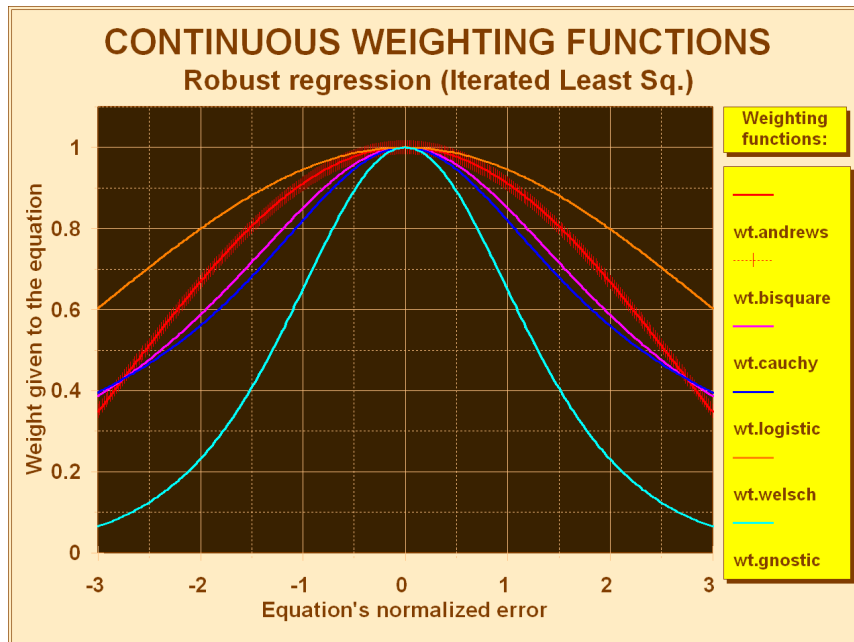


Figure 4: Continuous influence functions of robust statistics and of mathematical gnostics

An alternative weighting function based on the gnostic approach described below is also shown in Fig.4. Its form was theoretically proved and published in the paper⁴.

[1.4. Robust correlation methods available under S-PLUS](#)

Correlation coefficients play an important role in statistics especially in connection with linear relations between variables. Quality of estimates of correlation matrices is a limiting factor in tasks like regression problem, principal component analysis, factor analysis and discrimination analysis. They also can be applied to indicate that a certain similarity or even causal relation exists between variables or data samples. However, classical (Pearson's) estimates are unrobust because they are based on unrobust statistics (the first and second statistical moments). Moreover, their application and testing assumes the Gaussian distribution of data.

Robust statistical alternatives to classical correlations were found among nonparametric (distribution-free) methods. Two of such methods are available in S-PLUS system represented by the function *cor.test*. By choosing a parameter of this function, one can evaluate one of three versions of correlation coefficients and test the coefficient's value statistically:

- 1) Pearson's (classical) correlation coefficient,
- 2) Spearman's rank correlation coefficient,
- 3) Kendall's rank correlation coefficient.

Unlike the classical methods, the nonparametric methods can be applied without some rigid assumptions to data models and not only to metrical (measurable) quantities, but also to rank data subjected only to ordering.

⁴ Kovanic P.: A New Theoretical and Algorithmical Basis for Estimation, Identification and Control, *Automatica* **22**, No.6 (1986), 657-674



1.5. An alternative method available under S-PLUS: the mathematical Gnostics

The gnostic theory of uncertainty in individual data and small data samples was created in Czechoslovak Academy of Sciences and published in a series of three papers⁵ already in 1984. Presentation at the IX-th IFAC (International Federation of Automatic Control). Congress resulted in an invitation to publish the above cited paper in the official IFAC journal. A brief introduction into the approach follows.

Data uncertainty can be seen as a distance of the data item's observed value from the true one with the sign showing the direction of the deviation. Way of measuring distances is thus a matter of geometry.

To measure data uncertainty manifested by data errors, classical statistics applies the Euclidean geometry. Unlike this, the robust statistical theory (e.g. introducing some "influence functions" in its M-estimates mentioned in 1.5) gives different weights to elements of data errors in different points of the data space. This is equivalent to application of a geometry of the Riemannian type.

The problem of geometry is in Gnostic theory of **individual** uncertain data solved by the mathematical line of reasoning based on the first axiom, which assumes that the data item resulted from a proper measuring or counting procedure. ("Proper" means "consistent", satisfying the conditions formulated by the measurement theory developing already since von Helmholtz's time⁶). Unlike the measurement theory assuming an ideal case of no uncertainty, Gnostic model considers the uncertain data as points of a bi-dimensional plane, where the image of an observed data item is driven by its uncertain component along a path, the form of which is proved and shown to have extreme features of the Minkowskian circle. Nonlinear formulae of data error and data weight are attached to the points of the path. This path of virtual movement from the true to observed data value is the model of **quantification** (of the measuring or counting process). The way of optimum estimation of the true data value (the **estimation** path) is then found so to minimize the effect of the uncertainty. Models of quantification and estimation enable the formulae of entropy, information and probability of the individual data item to be derived. An equation of mutual entropy-to-information conversion is also proved.

To apply the theory of individual data to the **samples of uncertain data**, the problem of **uncertainty composition law** has to be solved. This is done by the second axiom. Unlike statistics, which composes data (and their products and squares) additively, additive composition of other nonlinear functions of data errors and weights is accepted. The formulae valid for the estimation process differ from those of the quantification. A different kind of robustness result for composed data. The estimating formulae ensure the robustness with respect to **outlying** data, while their quantifying versions are robust with respect to **inliers** such as inner disturbances or noise of the data sample⁷.

The formula of probability density of an individual data item can be viewed as a Parzen's kernel. However, the form of this kernel is uniquely determined by the theory. To compose individual uncertain data to get characteristics of a sample of uncertain data is mathematically proved to be of two complementary kinds:

- A) Additive composition of kernels producing the so-called **local** distribution function.
- B) Normalized additive composition of kernels resulting in **global** distribution function.

In dependence on the required kind of robustness and on the type of estimating or quantifying formulae applied, four distribution functions can be obtained:

The *estimating local* distribution function (ELDF) is universally applicable to all data samples. It is very flexible; it can fit arbitrarily spread data,⁸ by setting its suitable scale parameter. The choice of the scale parameter is dependent on the task. This is essentially a smoothing operation. Inevitably there is a trade-off between bias in the estimate and the estimate's variability: large scale parameters will produce smooth estimates that may hide local features of the density. It is similar to the decision on resolution power of an optical instrument: to observe stars, one needs a telescope, to see structure of a matter, a microscope is applied.

⁵ Kovanic P.: Gnostical Theory of Individual Data , Problems of Control and Information Theory 13 (1984), 4, 259-274.
Kovanic P.: Gnostical Theory of Small Samples of Real Data, Problems of Control and Information Theory 13 (1984), 5, 303-319. Kovanic P.: On Relations between Information and Physics, Problems of Control and Information Theory 13 (1984), 6, 383-399

⁶ Helmholtz H.von, Zaehlen und Messen erkenntniss-theoretisch betrachtet, in Philosophische Aufsätze Eduard Zeller gewidmet, Leipzig (1887), s,17-52.

⁷ An example: "normal" (good) products have sizes deviating from the norm only slightly and a quality assessment control does not undertake any actions with respect to them. However, it must react when a deviation exceeds the tolerance. Such a system should be robust with respect to inlying data and increasingly sensitive to the "outliers".

⁸ Data samples including both outliers and inliers as well as composed of several homogeneous subsamples.



The *estimating global* distribution function (EGDF) provides a **unique** estimate of the probability distribution function and its density in application to a homogeneous data sample because its three parameters (scale parameter and the lower (LB) and upper (UB) bounds of the data support) are estimated automatically under the condition of the best fit of the Empirical Distribution Function. This optimal estimation is robust with respect to outliers. The EGDF is unimodal when applied to a homogeneous data sample. Appearance of another mode signals the data sample's non-homogeneity. This feature enables reliable homogeneity test to be performed. To estimate the EGDF, only data are needed.

The *quantifying* versions (QLDF and QGDF) of both distribution functions differ from the estimating ones by their robustness: they suppress inliers and prefer peripheral data..

The choice of the kind and type of the distribution function depends on the formulation of the task: if a global view on the data sample, uniqueness of the distribution's parameters and testing the data's homogeneity is required, the EGDF is applied. To show the structure of a non-homogeneous data sample, ELDF is to be used with a scale parameter corresponding to the required resolution power of the analysis. If the robustness with respect to inliers is asked, the QGDF and QLDF can be applied. A preliminary version of the Gnostic software supported by the S-PLUS package along with a detailed Guide exists⁹. Its first (introductory) part is submitted as an Appendix to this report.

Gnostic solution of the regression problem is also available by using the Weighted Least Squares method mentioned above with the Gnostic version of the weighting function shown in Fig.4, the form of which was theoretically proved to be optimal.

The third class of tasks subjected to tests, i.e. correlation, also has its Gnostic solution worth to be mentioned in some detail: statistical definition of the correlation coefficient is based on deviations of the data vector's components from their mean. These represent data "errors" measured by application of the Euclidean geometry. Mean product of such error vectors are then normalized by products of their Euclidean lengths. Unlike this, Gnostic evaluation of an relative error (data uncertainty) of a true positive data value Z_0 observed as a positive Z applies a quantity h called *estimating irrelevance*:

$$h = \left(\frac{1/Q - Q}{1/Q + Q} \right) \text{ where } Q = \left(\frac{Z}{Z_0} \right)^{(2/S)}$$

and where S is a scale parameter. (Data, that can be negative or zero, must be transformed onto the interval of positive numbers before substitution into the formulae.) Irrelevance has thus values between -1 and +1 and equals zero for a precise data item Z . The *estimation weight* w is related with the irrelevance by the formula

$$w = \sqrt{1 - h^2}$$

while probability p of the data item to have value less or equal to Z is

$$p = (1 - h) / 2 .$$

Gnostic correlation coefficient is then determined in the same way like the statistical one but instead of the Euclidean data error, Gnostic errors $p - \frac{1}{2}$ are used. There also are quantifying versions of the irrelevance and data weight determined by formulae differing from the estimating ones. They ensure the robustness opposite to that of estimating versions.

A more detailed introduction to Gnostic theory can be found in the first part of the mentioned Guide, submitted as an Appendix of this report. Source programs of Gnostic functions implemented on S-PLUS reveal all details of the Gnostic algorithms.

Many examples of applications of the Gnostic methods exist resulting from the long-term experience. Specifically, solution of different problems of environmental data analysis can be found in the Guide along with examples of other application fields. Many examples of applications to the economical and financial problems have been published in books¹⁰ and were also addressed in the yet unpublished book¹¹.

⁹ Kovanic, P.: Guide to Gnostic Analysis of Uncertain Data, Institute of Public Health, Ostrava, Czech Rep., (2008), 439 pp.

¹⁰ Kovanicová D., Kovanic P.: Treasures hidden in accountancy (in Czech), Part II.: "Financial statement analysis", Polygon, Prague: 1-st edition (1995), ISBN 80-85967-07-03, 300 pp., 2-nd edition (1996), ISBN 80-85967-07-3, 300 pp., 3-rd edition (1997), ISBN 80-85967-56-1, 303 pp., 4-th edition (1999), ISBN 80-85967-88-X, 303 pp.



TESTS ON TYPICAL CASES MET IN RISK ASSESSMENTS

The selection of relevant input data and of their uncertainty can be classically difficult because of the structure of the dataset available. The following problems can indeed be met:

- the number of data is very low (typically less than 10) if monitoring constraints prevents the collection of large datasets;
- some data belonging to the dataset are under the level of detection (and can thus be considered as semi-quantitative/censored data);
- some peripheral data (outliers) can decrease the robustness of the analysis;
- the dataset is composed of well-monitored and badly-monitored pollutants respectively, and suspected correlations between both could improve the knowledge about the latter ones.

Some examples are presented for each of these cases, with a comparison of methods previously described.

1. Limited set of homogeneous data

To test and compare the probability density estimation methods, data from the Czech national monitoring of Permanent Organic Pollutants can be used, specifically, summary concentrations of PCDD/F shown in Tab.1.

Table 1: Concentrations of sums of PCDD/F in Czech and Moravian rivers (ng/l) in 2000

Code	Sum (PCDD/F)
4468	0.000688
4469	0.000610
4470	0.00253
4471	0.0183
4472	0.1052
4473	0.0228
4474	0.00503

Locations of the measurements were identified by the code in the first column of the table.

[Application of the statistical kernel estimation methods](#)

Kernel estimates of the probability density functions obtained by application of four statistical methods described in 1.4 are shown in Fig.5 for the default values of the parameter *width*. Concentrations are positive numbers but kernels applied to their measured values partly reach the negative interval. A more realistic representation can be obtained by application of the kernel method to logarithmic data values, as was demonstrated in Fig.2.

Kovanicová D., Kovanic P.: Treasures hidden in accountancy (in Czech), Part III.: "Financial control of the growth rate of a firm", Polygon, Prague, 1-st edition (1996), ISBN 80-85967-35-9, 280 pp., 2-nd edition (1997), ISBN 80-85967-58-8, 280 pp.

¹¹ Kovanic P., Humber M.B.: The Economics of Information, Mathematical Gnostics for Data Analysis, unpublished, 728 pp.

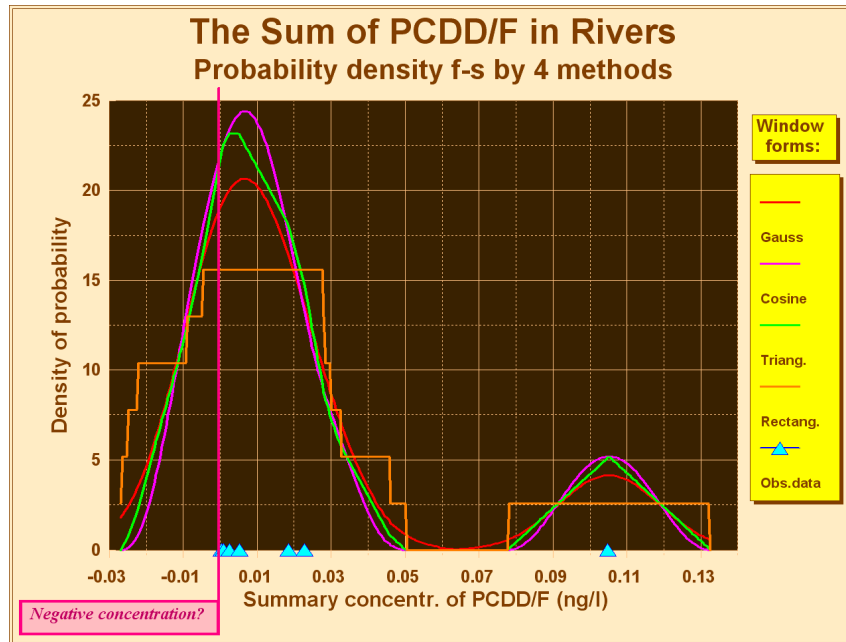


Figure 5: Statistical estimates of probability density functions

Gnostic global probability and density function

Gnostic global and local distribution functions (EGDF and ELDF) are in Fig.6 and Fig.7.

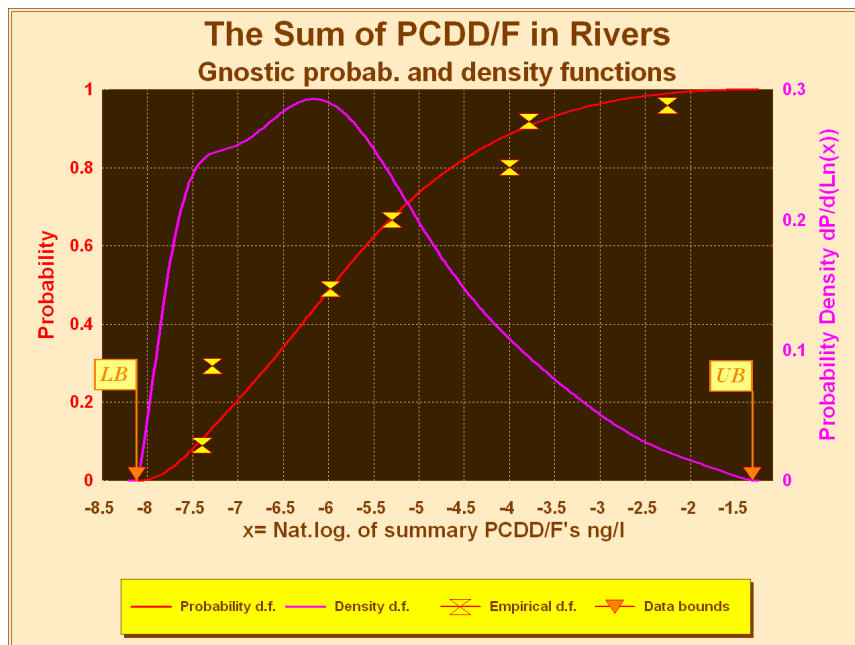


Figure 6: Global distribution function EGDF and its density of the summary PCDD/F

The same data were used, but instead of their values, the middle points of the steps of the Empirical Distribution Function (EDF) are shown by yellow marks because the EGDF is optimized to be the best fit of this EDF. After



estimation of the both bounds and of the scale parameter, the algorithm computes the probability density function defined as logarithmic¹² derivative of the probability function.

Bounds of the data support LB and UB estimated by means of the EGDF's algorithm are also shown.

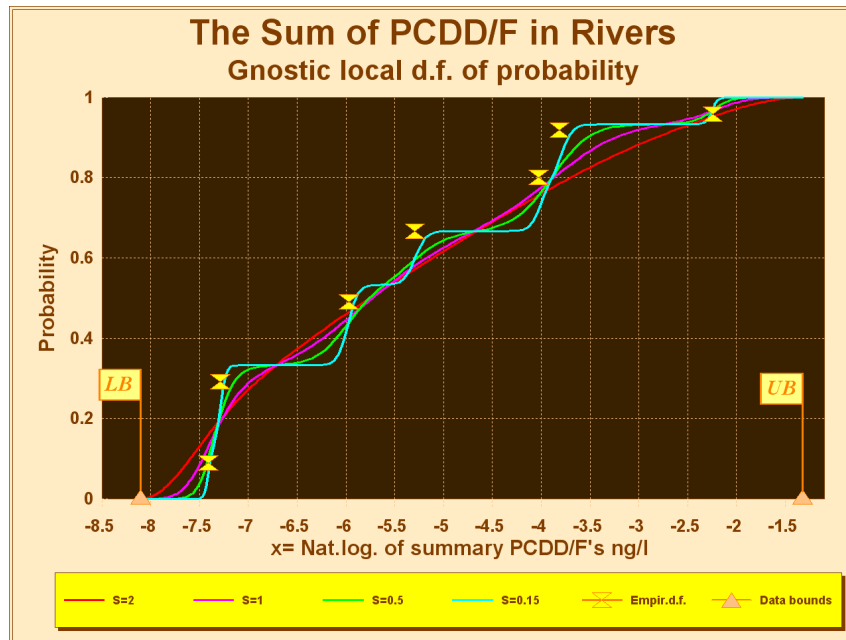


Figure 7: Local distributions (ELDF) with different scale parameter (S).

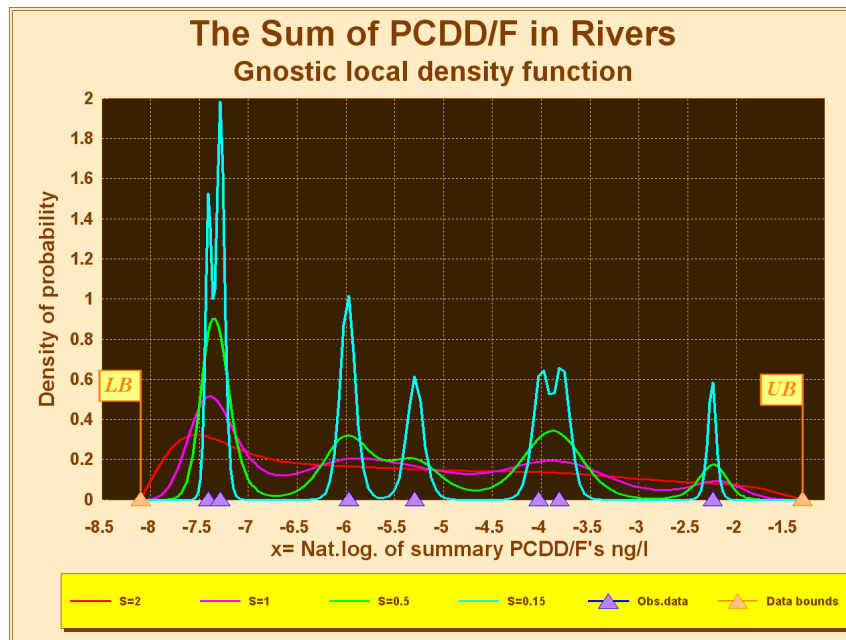


Figure 8: Distributions of the probability density (ELDF) with different scale parameter (S).

It is seen in Fig.6 that the EGDF is rigid enough to not adapt its form to individual data deviating from the smooth data fit. A filtering effect onto the data that do not follow the general form of the distribution is

¹² Logarithmic derivative is used to ensure consistency of the density form with the logarithmic x-axis applied.



exhibited. Unlike this, the local distribution ELDF's flexibility is controlled by the scale parameter's value to reveal all required details of the data sample's structure. Decreasing of the parameter S results in diminishing the vertical distances of the probability function from the yellow marks in Fig.8 and making the "Gnostic kernels" in Fig.8 narrower.

2. Dataset containing semi-quantitative data (lower than the detection limit)

Data applied for the test of treatment of the data measured below the Limit of Detection (LOD) (the *low-censored data*) originated in a survey made by the team of the National Institute of Public Health, Prague, Charles University, Prague and Institute of Public Health, Ostrava. Analysis of results were published¹³.

Eighty measuring of the 2378-TCDD within this study resulted in 17 'properly' measured values while 63 results appeared to be below the LOD. Results of the test are presented in Fig.9 in the form of four EGDFs computed under the accepted assumptions.

The goal of the test is to compare four methods of treating this mixture of 'proper' and low-censored data:

- 1) To take into account only the 'proper' data and to ignore the censored ones;
- 2) To assume that the censored data had values equal to the LOD;
- 3) To assume that the censored data had values equal to the LOD/2;
- 4) To apply the gnostic EGDF capable to make use of both 'proper' and censored data.

Probability density functions corresponding to Fig.9 are in Fig.10.

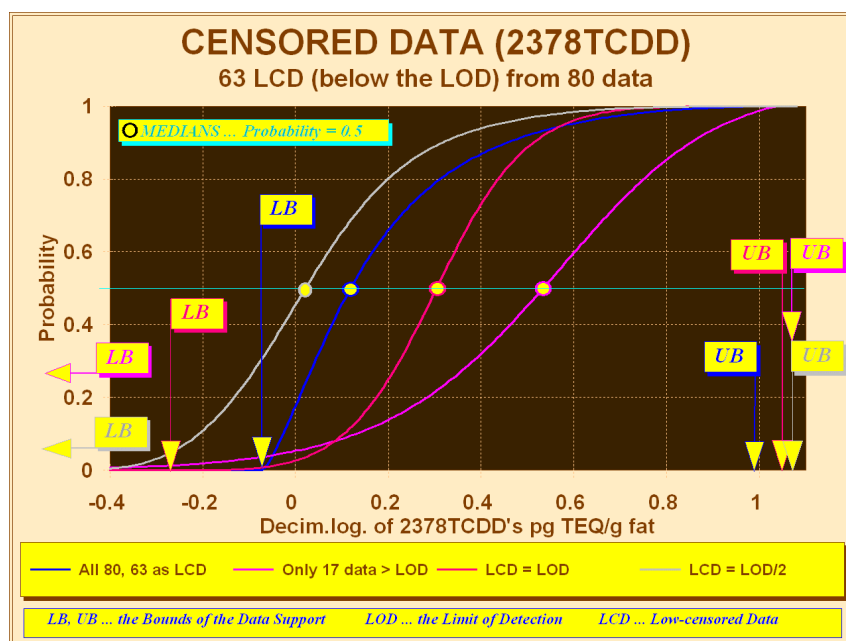


Figure 9: Probability distributions of results of four approaches to low-censored data

¹³ Černá, M. et al., Levels of PCDDs, PCDFs, and PCBs in the blood of the of the non-occupationally exposed residents living in the vicinity of a chemical plant in the Czech Republic, Chemosphere, Vol 67, Issue 9, 238-246(2007), doi:10.1016/j.chemosphere.2006.05.104

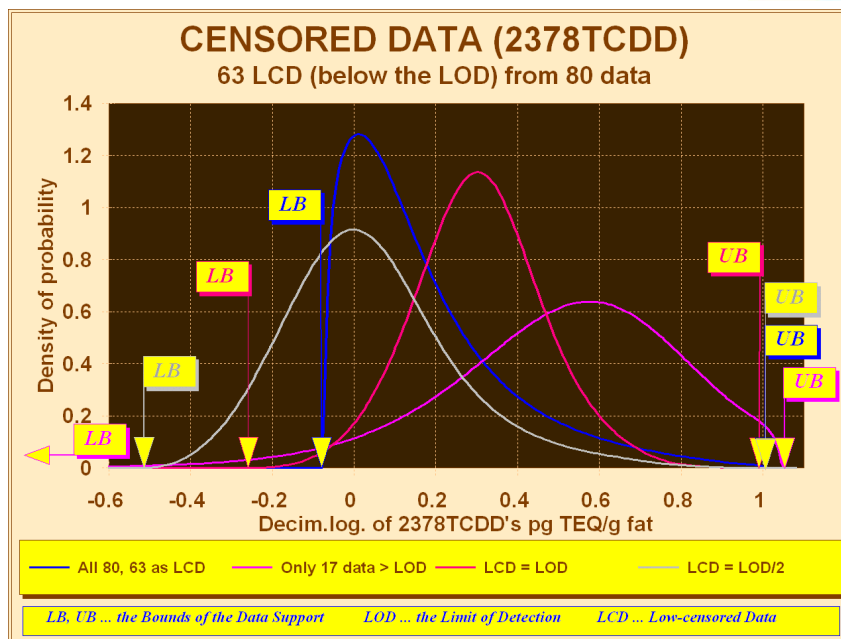


Figure 10: Density distributions of results of four approaches to low-censored data

3. Dataset containing outliers or up-censored data

The fundamental problem of outliers is their recognition: what value has the bound of the “proper” data, exceeding of which by a data sample’s item is to be considered as outlying?

In the most fundamental case of occurrence of such a problem, in the theory of mathematical sets, the notion of the *membership problem* is applied: is an object x a “member” of the set X in the sense of $x \in X$? The classical set theory considers this problem as “primitive” one, i.e. “everyone knows if x belongs to X or not”. Unlike this, the fuzzy set theory introduces the notion of the membership function to quantify a degree of membership by a number. Another approach is used in statistics, e.g. in quality assessment practice: the size (or other “normed” quantitative characteristic of a product) is an outlier if it deviates from the mean value more than K -multiple of the standard deviation of the “normal” products. All three mentioned approaches are based more on subjective judgement than on facts (data). Moreover, mean values and standard deviations have a good sense only for some special probability distributions of data to attach the probability to the data value.

Unlike this, the membership problem is solved uniquely in Gnostics by using the notion of

data sample’s homogeneity¹⁴: a data item is an outlier if its inclusion into the homogeneous sub-sample of the data makes this sub-sample non-homogeneous.

An example can illustrate the problem (Fig.11):

Concentrations of emissions of pollutants were monitored in a Czech city in 1999. There are three bounds shown in Fig.11 enabling to classify the data as “normal” or “outliers”. (“Normal” means “usual”, not Gaussian in this connection.) The Gnostic bound determined 5 outliers while 2STD were exceeded four times and the frequently used deviation 3STD was exceeded only once. Consequences of the subjectivity of setting the multiplicity of STD are obvious here. Unlike this, the value of USB was obtained objectively and uniquely, by data analysis. This corresponds to the idea “Let data speak for themselves”.

¹⁴ A data sample is considered homogeneous in Gnostics when its probability density of the “bell” form has only one maximum or if its density of the U-form has two maxima. This test is robust due to the robustness of the EGDF.

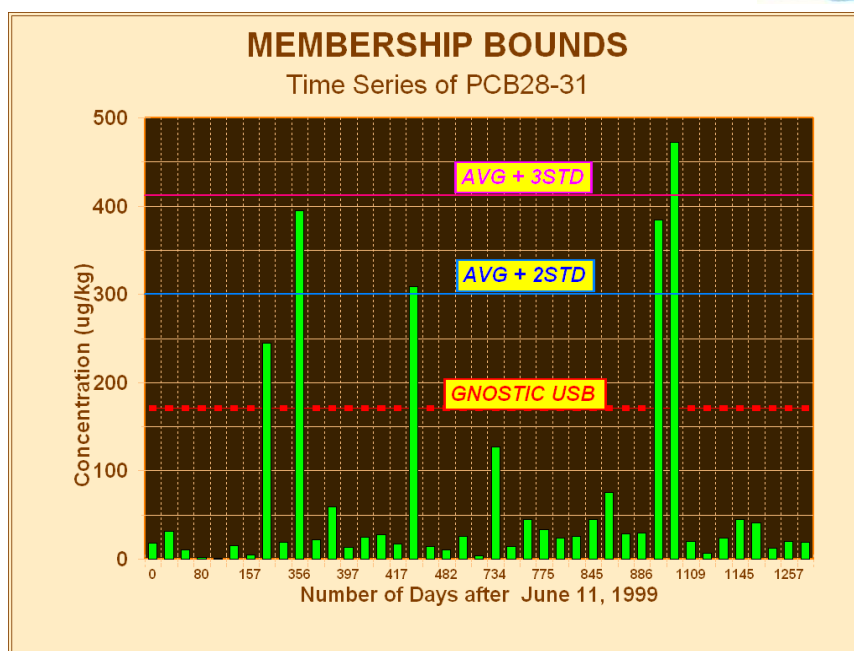


Figure 11: Comparison of the Gnostic upper membership bound USB with two statistical ones

A remark is worth to be added on the up-censored data. These are defined as exceeding the LOR (Limit of Range). It is easy to see that they might be outliers as well as the “proper” data: try to measure the voltage at home (e.g.220V) by a voltmeter the range of which is only 100V. The measured value (full scale) will only say that the voltage in the plug exceeds 100V. However, such information can be of fundamental importance in environmental monitoring. The significance of proper treatment of up-censored data lies its capability to determine *that* the LOR was exceeded but also to estimate *how far* the quantity reached.

4. Tests of regression methods

Data for the MD-test also originated in the survey cited in connection with the 2378TCDD.

In 2003, concentrations of altogether 17 PCDD/Fs congeners and 12 non-ortho and mono-ortho dioxin-like PCBs were measured in the blood of 60 randomly selected adults who lived in three settlements surrounding a chemical plant that had been producing chlorinated herbicides (mainly HCHs, HCB, pentachlorophenole, 2,4,5-T) in the 1960's; subjects consuming home-produced animal foods were chosen. Twenty blood donors with similar characteristics from the locality with about 56 km distance were used as control subjects. The factors that influenced the dioxin levels were investigated on the basis of a questionnaire. The aim of the study was to find out whether the residents living in the surroundings of the chemical plant were at a greater exposure risk than the controls. To calculate TEQ values, WHO-TEFs were used. The concentrations of four PCDD and six PCDF congeners were below the LOD in more than 50% of samples.

To reach the goal of the survey, classical statistics such as means, medians and standard deviations were determined in the cited report. Instead, robust multi-dimensional regression models will be applied below to analyze impact of some quantitatively determined variables onto the accumulation of POPs in people. The dependent variables will be the total value of 29 pollutants found in blood of the individual persons quantified in pg TEQ/g fat. This variable will be denoted TotPOP. The explanatory variables will be the following:

- Constant: “Intercept” (the part unexplained by the model).
- x1: “Age” (years).
- x2: “Distance” (distance between the chemical plant and permanent residence in km).
- x3: “Alcohol” (grams of 100% alcohol consumed in wine or beer per a week).
- x4: “HealthSt” (sum of number of diagnoses and medicaments permanently taken).
- x5: “BMI” (Body Mass Index in kg/m²).
- x6: “Smoking” (number of cigarettes smoked per a week).

The regression model will include the explanatory variables linearly; only the distance will influence the TotPOP in a quadratic way, because this results in a better model than the linear form.



All the eleven robust methods with their weighting functions depicted in Fig.3 and Fig.4 were tested in application to these MD-data along with the gnostic weighting and with the classical (OLS, Ordinary Least Squares) method, which can be considered as a special case applying fixed weights all equaling one.

Results of the test are summarized in Table 2.

Quality of a multi-dimensional model is used to be quantified by the R-squared. It is a ratio of the variance explained by the model and total variance. Square root of the R-squared is known as the “multi-dimensional correlation coefficient” or as “coefficient of determination”. The last notion is applied in Tab.2.

Table 2: Results of tests of robust regression methods

Weighting function	Coef. of determination	Weighting function	Coef. of determination
OLS (all weights 1)	0.701	wt.huber	0.715
wt.andrews	FAILED	wt.logistic	0.714
wt.bisquare	FAILED	wt.median	FAILED
wt.cauchy	0.716	wt.talworth	FAILED
wt.fair	0.824	wt.welsch	0.704
wt.hampel	0.705	wt.gnostic	0.930

5. Tests of correlation coefficients

Data for the comparison of the methods originated in the national monitoring program of the Czech Republic, of which is the Institute of Public Health, Ostrava the certified reference laboratory. Fourteen Permanent Organic Pollutants and four inorganic ones were measured and correlations of their concentrations in the environment analyzed.

Values of correlation coefficients of mercury with the other pollutants evaluated by four methods mentioned above are shown in Fig.12 ordered by the unrobust (Pearson’s) results marked by red color.

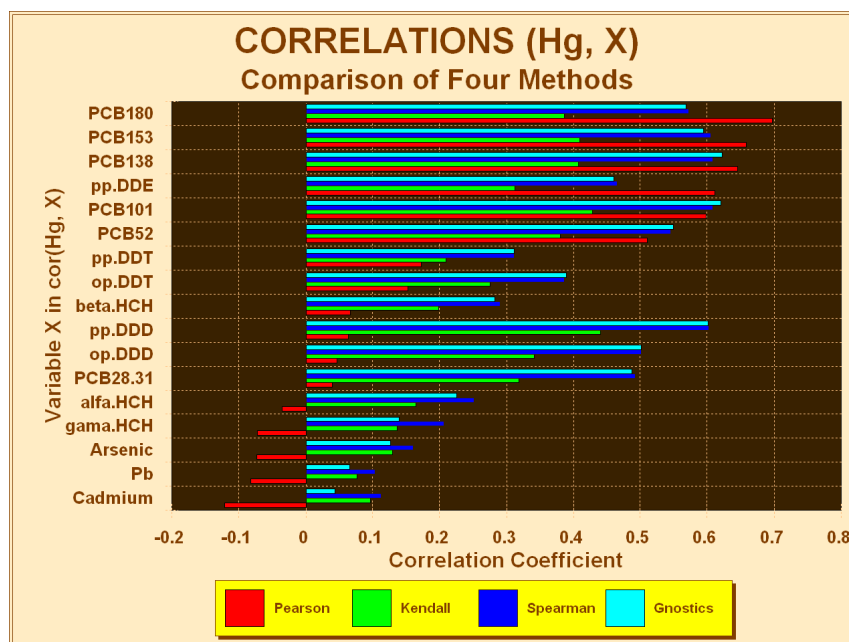


Figure 12: Comparison of four methods for estimation of correlation coefficients



DISCUSSION

1. Probability and density distributions

Statistical density estimates using basic kernels

Statistical kernel estimates using the four types of kernel functions programmed in S-PLUS obtained by the function *density(...)* were shown in Fig.5 for the default value of the parameter *width*. All four forms of windows/kernels satisfy the requirement of convergence with unlimitedly increasing number of data but in application to finite data samples, the resulting density function is not smooth and differentiable in three cases from four because of the non-smooth form of the kernels. However, neither the Gauss kernel can be applied to the treated positive data directly because the resulting density function is positive over the unrealistic negative part of the data support, although concentrations are always non-negative. By application of the function *density* to data subjected to natural logarithmization, Fig.2 is obtained. The continuous function applying the Gaussian form of the kernels is red. It has three maxima when the default value of the parameter *width* is used. The effect of different values of this parameter was demonstrated in Fig.1.

It can be seen from the examples that there is much subjectivity in the described way of estimating the probability density functions by means of statistical kernels:

- 1) There is no theoretical proof, which kernel is the optimal choice from the infinite number of continuous and smooth kernels satisfying the Parzen's convergence conditions;
- 2) Some of "proper" kernels make the resulting density function non-differentiable;
- 3) In case of outlier(s), the resulting density function can unrealistically (and unrobustly) extend the domain of other data;
- 4) There is no recommendation to the optimality of the parameter *width* desirable for a global representation of the data structure;
- 5) Data bounds (bounds of the data support, of the domain of non-zero density of probability) is not estimated using the data but subjectively given by the user (by his selection of the kernel's form and width or by his setting of the procedure's parameters *from* and *to*);
- 6) There are tasks for which the (cumulative) probability distribution is needed. To obtain this function, integrals of kernels must be composed;
- 7) There are no tools provided to a direct evaluation of the quality of fitting the true data distribution;
- 8) It is not obvious how to define a location parameter of the data sample;
- 9) Parameter *width* is set by the user in an experimental way to satisfy requirements to the detailedness of the insight into the data structure but there is no unique criterion to decide if the data sample is homogeneous or not.

Gnostic probability and density distribution functions

Application of the *global* distribution function (EGDF) to the same data as in previous examples is demonstrated in Fig.6.

Both probability and density functions are estimated simultaneously because both are needed for different applications, for example in testing some hypotheses. The discrete marks in Fig.6 are a direct representation of the treated data. They are placed in the middle points of the steps of the EDF (Empirical Distribution Function) used in statistics. It is known that these points are used by the Kolmogorov-Smirnov statistical test of fitting the data by a distribution function. In the case of the EGDF, distribution function's parameters LB, UB and S are automatically estimated so to minimize distances between the continuous distribution and EDF. The robustly estimated bounds LB and UB are also shown in the figure. They offer a fundamental information on the observed process: concentrations lower than LB and exceeding UB are unexpected, estimated probabilities of such events are zero.



To demonstrate features of the *local* distribution function ELDF, Fig.7 and Fig.8 are applied. Unlike the rigidity of the EGDF resulting in an objective and unique smooth data fit, the local distribution function ELDF allows to make a compromise between a smooth data representation and revealing the details of the data structure. A small scale parameter S decreases the distance between the ELDF and EDF in Fig.7 and makes the relief of local peaks of the probability density in Fig.8 more plastic. This may remind of the effect of the parameter *width* in Fig.1. However, there is a substantial difference between the forms of kernels, because that of the ELDF is theoretically proved as the “natural” one resulting from the nature of the uncertainty. This form ensures a smooth representation of the data distribution due to the smooth link-up of individual kernels.

One of possible aspects of the choice of the scale parameter is the number of peaks of the density model. One local maximum is shown with $S=2$ (red line), four with $S=1$ (magenta line), five with $S=0.5$ (green line) and seven with $S=0.15$ (cyan line). Some edges can be seen on the cyan line, but this is only due to insufficient number of the drawing points (400), the density function remains to be smooth and differentiable even with small values of S .

Goodness-of-fit of all Gnostic distribution functions is quantified by evaluation of the distances of the smooth representation from the middle points of the EDF. There also is a Gnostic evaluation of the fit: the relative amount of information borne by the distribution function (an information efficiency of the data treatment measured in per cents of the all information borne by the input data). This result along with estimated S , LB , UB , location and other numeric parameters and distribution graphs form the output of the Gnostic functions estimating the distributions.

There also is a disadvantage of the approach: demandingness with respect to time of computing caused by the necessity of optimization within a three-dimensional space of S , LB and UB . However, progress of computers gradually diminishes this flaw that is compensated by maximization of information obtained.

2. Estimation of distribution functions of incompletely defined (censored) data

Realization of the classical kernel estimation method in S-PLUS does not offer a way of incorporating the censored data in the analysis. However, such data (especially those measured below the sensitivity threshold of the measuring) play an important role in environmental control, because (a) they are the most desirable ones and (b) even a low level concentration of a pollutant can become dangerous due to long-term accumulation in organisms. Necessity of taking such low concentrations in account brought analysts to application of several substitutive methods cited in 2.2. Gnostic distribution function EGDF enables making use of information borne by the censored data (low- and up-censored ones and interval data) owing to its robust estimation of the data support bounds. Comparison of results of the three indirect methods with the Gnostic one based on probability distribution functions in Fig.9 and their densities in Fig.10 documents differences between the approaches of orders of magnitude.

The urgency of this problem can be confirmed by a citation¹⁵ on its another recent “substitutive” solution: to estimate the concentration of the unmeasurable TCDD as 40% of measurable PeCDD. This ratio was found as average for the general population. However, this ratio cannot be accepted as a universal constant. Other values can be derived from other surveys.

3. Robust regression models

As results from Tab.2, methods *wt.andrews*, *wt.bisquare* as well as *wt.median* and *wt.talworth* failed in application to considered data. This is due to the finite domain of the weighting functions shown in Fig.3: a large equation error exceeding a domain’s bound causes a zero equation weight that is not accepted by the function minimizing the weighted least error squares.

Methods *wt.welsch* and *wt.hampel* appeared to be only slightly more robust than the unrobust method Ordinary Least Squares. The rest of methods demonstrates usefulness of application of robust methods, especially of the gnostic one that gave the best result.

¹⁵ Needham L.L. and all.: Assigning concentration values for dioxin and furan congeners in human serum when measurements are below limits of detection, *Chemosphere* 67 (2007) 439-447.



Mean impacts on the sum of POPs found in blood of tested people can be evaluated using the regression models. Such values are shown in Fig.13 for the best two methods (Fair and Gnostic) along with the results of the unrobust classical OLS.

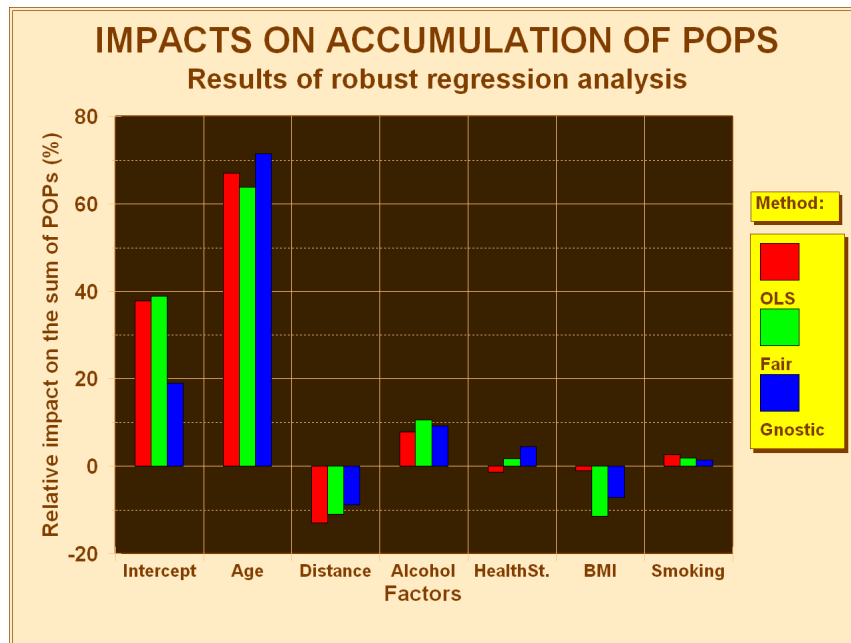


Fig.13: Results achieved by two robust methods in comparison with the unrobust one

Following conclusions can be inferred from the results:

- 1) Application of the unrobust OLS method cannot be recommended:
 - A) It yields obviously mistaken sign of the relation between the health status and accumulation of pollutants: “the more POPs in blood, the better health”,
 - B) The impact of BMI is obviously underestimated, because increasing personal weight surely results in decreased concentration of pollutants.
- 2) The most important impact exceeding 60% is logically that of the age, because amount of accumulated pollutions increases with the exposure time.
- 3) Results of different methods can substantially differ.
- 4) The least unexplained component of the amount of POPs in organisms and the best value of the coefficient of determination resulted by the Gnostic method.

The robust statistical methods available in S-PLUS were theoretically derived for some classes of data under some assumptions on their features. However, the ideal is to have methods not only robust to some ‘unusual’ data but also to assumptions on data. It is unrealistic to think of a completely assumption-free method, but the assumptions should be as simple as possible and the class of ‘suitable’ data as broad as possible. Validity of Gnostic methods is limited by the requirement of the first Gnostic axiom: data must have simple algebraic features of the Abel’s additive or multiplicative group to be consistent images of real quantities in the sense of the measurement theory. This is what can be simply defined as ‘real data’. Such a category is very broad.



4. Correlation coefficients

Tests' results of four methods depicted in Fig.11 enable following observations:

1. Classical (unrobust) results tend to overestimate strong correlations while failing in cases of weak correlations.
2. Robust statistical estimates (Kendall's and Spearman's) strongly differ from the classical ones, but they also substantially differ each from the others.
3. Gnostic estimates are in 10 cases from 17 between Kendall's and Spearman's while being in 14 cases from 17 closer to Spearman's than to Kendall's ones.

Differences between results of four methods can be quantified by calculating mean squared differences D between vectors' components. These are shown in Table 3.

Table 3.: Comparison of estimating methods for correlations

i	j	Method i	Method j	D
1	2	Pearson	Kendall	0.049
1	3	Pearson	Spearman	0.173
1	4	Pearson	Gnostics	0.171
2	3	Kendall	Spearman	0.124
2	4	Kendall	Gnostics	0.122
3	4	Spearman	Gnostics	0.002

Closeness of the Gnostic results to Spearman's are thus confirmed in spite of the quite different theoretical fundamentals of both approaches. Another factor supports the validity of the Gnostic approach to correlations: it is well-known that an important role in estimating the parameters of regression models is played by the correlation matrix of the explanatory variables. As shown in Tab.2, the weighting functions based on the Gnostic theory lead to the best solution of the regression problem.

CONCLUSIONS

Tests of selected methods for the treatment of data originated from monitoring programs were performed. Both classical and robust statistical methods were tested and compared with an alternative (Gnostic) method. All the selected methods are available as functions supported by the American software package S-PLUS. Four tasks, solution of which is frequently needed in the environmental control, were subjected to tests:

- 1) Estimation of probability and probability density distribution functions of a small data sample;
- 2) Estimation of true data values in case of measuring a part of data below the detection limit by means of a robust probability distribution function;
- 3) Estimation of parameters of a multi-dimensional model of impacts of living conditions of peoples on their accumulation of pollutants;
- 4) Estimation of correlation coefficients between pollutants.

Summarizing the discussion to all four classes of tests (3.1 – 3.4), it can be concluded that Gnostic methodology is suitable for application to considered problems. It proves useful in robust yielding more information from small samples of uncertain data than the tested commercially available statistical methods. Other methods, such as fuzzy logic or geostatistical methods exist and might be of use, but fall outside the scope of this document.