



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**

– EXTENDED REPORT ON REVIEW AND TESTS OF METHODS FOR FILLING DATA GAPS –

Due date of delivery: *31/07/2009*

Actual submission date: *28/08/2009*

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 Extended report on review and tests of methods for filling data gaps
ID D1.9.doc
Revision Final version
Revision Date 09/09/2009
Authors P. KOVANIC (IPH) and T.OCELKA (IPH)

Approvals

	Name	Company	Date	Visa
Author	P. KOVANIC	IPH	20/07/2009	P. Kovanic
Co-Author	T. OCELKA	IPH	20/07/2009	T.Ocelka
WP Leader	A. MARCOMINI	UNIVE	28/08/2009	po E. Giubilato
Coordinator	F. BOIS	INERIS	09/09/2009	F. Bois

Documents history

Revision	Date	Modification	Author
Version 0	18/01/2008	Template made available to the WP leaders	F. BOIS
Version 1	20/07/2009	First version	P. KOVANIC
Version 2	09/09/2009	Final version	F. BOIS



Contents

INTRODUCTION	5
1. THEORETICAL ASPECTS OF AN APPROACH TO A MODEL OF UNCERTAINTY	6
1.1. The data model	6
1.2 The uncertainty model	6
1.3 The problem of geometry	7
1.4 The problem of entropy	8
1.5 The problem of information	8
1.6 The problem of additivity	9
1.7 The problem of optimality	9
1.8 Choosing the data treatment method	10
2. RESULTS OF THREE CASE STUDIES	11
2.1 Problems of testing a data treatment method	11
2.2 Used databases	11
2.2.1 Preliminary database	11
2.2.2 Database from monitoring the rivers of the Czech Republic	12
2.2.3 Database from monitoring the groundwater contamination in Poland	12
2.3 Analysis of bio-monitoring data	13
2.3.1 Effect of left-censored data	13
2.3.2 Robust testing of hypotheses	13
2.4 Analysis of the surface waters	17
2.4.1 Correlations in pollutants and toxicities	17
2.4.2 Robust models of interactions in pollutants and toxicities	17
2.5 Analysis of the groundwater contamination	19
2.5.1 Marginal cluster analysis	20
2.5.2 Homogenization of the sub-samples	21
2.5.3 Background and total contamination	21
2.5.4 The impact of the river on the spread of contamination	22
2.5.5 Dynamics of contamination	22
2.5.6 Interdependence of observed variable	24
2.5.7 "Scientific water witching"	25
2.5.8 Conclusions of the environmental case studies	26
3. LONG-STANDING EXPERIENCE WITH GNOSTIC METHODS	27
3.1 Applications in economics	27
3.2 Applications in technology	27
3.3 Applications in monitoring of environment	28
3.4 Other applications	28



4. COMPARISONS OF ROBUST METHODS	29
4.1 Comparison of robust location estimators	29
4.2 Comparison of non-parametric distributions	31
4.3 Comparison of robust estimates of correlations	34
4.4 Comparison of robust regression models	36
5. CONCLUSIONS	40
6. ACKNOWLEDGEMENTS	41



INTRODUCTION

Science does not develop all the time “linearly” (smoothly, by small improvements) but its significant progress is realized by “scientific revolutions” consisting of changes of paradigms¹. The 2-fun project has the uncertainty in its definition. It is important to take notice, that uncertainty has been an unending problem of science for centuries. As its “twin” - risk it deserved a recognition of historical milestone: “The revolutionary idea that defines the boundary between modern times and the past is the mastery of risk...”². Attempts of mathematicians to manage the risk can be traced to time of Luca Paccioli (1445-1514/1517), who brought double-entry bookkeeping to the attention of the business managers of his day and tutored Leonardo da Vinci in the multiplication tables. What followed can be interpreted as development of statistical paradigm of uncertainty, which was gathering great successes in many fields of practice including the science and technology. This paradigm became dominating in educating and teaching the students and in using for data treatment. This does not mean, that its leading role remains to be sacrosanct forever. There were sources of questioning this role in both theory and practice.

Risks are measured by probability. The original conception of statistics based on probability was based on the observation, that under stationary conditions relative frequency of occurrence of some events converges to a limit (“probability”). However, this statistical paradigm is not unique. There are altogether seven classes of theories of probability based on different paradigms which already were described in detail and analyzed³. The conclusions drawn therein are far from optimistic:

*“...The many difficulties encountered in attempts to understand and apply present-day theories of probability suggest the need for a new perspective. Conceivably, probability is not possible. A careful sifting of our intuitive expectations and requirements for a theory of probability might reveal that they are illusory or even logically inconsistent. Perhaps the Gordian knot, whose strands we have been examining, is best cut. However, where would such a drastic step leave the world of practice?
... Clearly much remains to be understood about random phenomena before technology and science can be soundly and rapidly advanced. It is not only the “laws” of today that may be in error but also our whole conception of the formation and meaning of laws.”*

Perhaps this sad state of affairs can be interpreted as a call for a good non-statistical paradigm to assess uncertainty. Need for such a change has been becoming generally accepted by the scientific community. This can be documented by the existence of the conference IPMU (Information Processing and Management of Uncertainty in Knowledge-based Systems), of which its 12-th meeting took place in 2008⁴ in Málaga, Spain. Topics of this meeting included 28 theoretic approaches (paradigms) and 24 fields of applications. These efforts can be interpreted as a running international competition of paradigms bidding for recognition.

There also are pragmatic reasons for looking for an alternative of statistics. The post-WW2 development of computers enabled mass applications of statistical methods to be realized and tested. The experience was far from satisfactory. This experience also motivated not only the development toward alternatives but also toward making the statistical methods more robust. A large number of different robust statistical approaches was created. The lack of robust methods has been exchanged by a superfluity of methods, of which some work in application to some data while failing with other data. Large scale testing of robust statistical methods has shown that to each data treatment task a suitable method and a number of non-suitable methods exist, but a new problem appeared: how to find the suitable method for application to some given data.

It is therefore obvious, that the task formulated as “Extended report on review and tests of methods for filling data gaps” must deal with both theoretical and especially pragmatic aspects to make the problems narrower and manageable. Moreover, there is a candidate available for a favorable role in the “competition of uncertainty paradigms” missing in the scope of the IPMU-conferences: the gnostic theory of individual uncertain data and small samples with its extensive basis of robust procedures maximizing the information of data analysis. This approach is to be included into the review and compared with the other approaches using both theoretical arguments and results of applications to real data.

¹ Kuhn T.S., The Structure of Scientific Revolutions, University of Chicago Press (1962)

² Bernstein P.L., Against the Gods, the Remarkable Story of Risk, John Wiley & Sons, Inc., New York (1996)

³ Fine Terrence L., Theories of Probability; an Examination of Foundations, Academic Press, New York and London, (1973).

⁴ <http://www.gimac.uma.es/ipmu08/topics.html>



Accepting the pragmatic aspect of the task and wishing to subject the selected methods to comparison in applications, it is necessary to limit the choice of methods, that are already available in the form of functions running on computers. Academic ideas published in journals neither congress presentations can hardly be tested without additional efforts resulting in usable software. Experience shows, that this aspect acts as an efficient “filter” making only two classes of approaches suitable for a detailed review and comparison: methods of robust statistics and gnostic methods. Nevertheless, criteria applied in comparison of these two classes might be applied to other alternative methods to judge on their viability.

1. THEORETICAL ASPECTS OF AN APPROACH TO A MODEL OF QUANTITATIVE UNCERTAINTY

1.1 The data model

The task is to evaluate the uncertainty of quantitative recognition of qualitatively delimited features of objects and processes of the real world. “To evaluate” means to map some structures of real quantities onto structures of numbers. Such a mapping is unusual for mathematics dealing with objects defined and manipulated within mathematics, while the discussed task maps the reality into mathematics. The development of science arrived at description of such an “unusual” mapping by using mathematical language and requirements of consistency already in 19th century⁵ by starting the development of the measurement theory. This theory bestowed the real measuring instruments (scales, meters, etc.) by an indispensable role in the quantification process. To be suitable for such a role, these must satisfy to requirements formulated in strict mathematical terms. These rules are not only submitted to the requirements of consistency of quantification process. They correspond to ideals of instruments’ serving to praxis. Moreover, these features can be practically tested. This makes the measurement theory not only theoretical, but also practical base for quantification producing the special product, *data*. From this point of view, data are not some arbitrary numbers, but images of real quantities passed to mathematics by the quantification process, which impresses to data some special features. The first requirement to the data treatment approaches is related to the data model: how these special data nature are respected.

There is a feature of the real quantities, which must not be omitted – their finiteness. Data as images of the real quantities cannot be infinite. However, many data models consider data domain as infinite intervals. Moreover, bounds of the data domain/support can frequently be the most desirable results of the analysis. A realistic data model must enable the data bounds to be estimated, but this is impossible in cases of bounds, which are infinite by a priori assumptions.

Many statistical methods start with the requirement: “Let the data be...” followed by the wish of the method’s author related to data features. However, the data features should result from the analysis of the data and are not available before the analysis. A priori assumptions on data are subjective. They represent an attempt to violate the image of the nature of real structures depicted by data. The ideal is “Let data speak for themselves”. Many of methods are far from this ideal.

The gnostic approach is based on an extension and generalization of the “unary” theory of measurements (dealing only with data not disturbed by uncertainty) resulting in a “binary” model of uncertain data. Both unary “threads” (the quantification channels) of its data model are conformed to the classical theory of measurements. The first and decisive axiom of the gnostic theory represents thus two applications of the measurement theory.

1.2 The uncertainty model

The uncertainty should be also accepted as something objectively existing, real. When the voltage of a source is to be measured, its observed value can be uncertain because of contributions of some undesirable changing electromagnetic fields. But these are real and their impact is of the same nature like that of the “desired”, “true” quantity. The ordinarily used a priori data models include both “true” and “uncertain” data components. The adjective “subjective” thus relates to the uncertainty model as well. The model of uncertainty is frequently more complicated than that of the “true” component, it can include assumptions on probability distribution function or

⁵ Helmholtz H. von, *Zahlen und Messen erkenntniss-theoretisch betrachtet*, in *Philosophische Aufsätze Eduard Zeller gewidmet*, Leipzig (1887), s.17-52



– at least – on the family of distributions and assumption of independence or non-correlations. All this can as well as cannot reflect the true nature of the structures of depicted quantities.

The comfortable and favorite “normal” (Gaussian) statistical model and its derivative (like chi-square, Student’s and Fisher’s distributions) are frequently used in spite of their theoretical and practical limitations to “sufficiently large” and “sufficiently good” data with small disturbances. Even a small contribution of strong uncertainty can cause distortion of results of methods based on such a priori models.

Statisticians are proud of being mathematicians. However, as such, they should not forget about the fundamental conditions, under which the Central Limit Theorem holds:

- 1) The random variable has some distribution with a mean and a standard deviation.
- 2) The data samples are randomly selected from a population.

However, means and standard deviations are integrals. How to be sure of convergence of such integrals when observing uncertain data with an unknown distribution? How to be sure of uncertainty randomness (“values obtained by chance from a population”) when only some particular data are available with no information on a population?

Gnostic model of uncertainty introduced by the triple of papers⁶ considers the uncertainties as some elements of a structure of real (unknown) quantities of the same nature as that of the true quantities. Model of the quantification is thus a two-dimensional mapping, but the observed data are one-dimensional projections of both components. Structures of quantities are taking as commutative groups. Their numeric images (data) are therefore commutative groups as well. The structure operation on the two-dimensional structure is identical with operations on both its components because of identity of their nature. This model accepted as an axiom allows many consequences of fundamental importance to be derived.

1.3 The problem of geometry

Error of an observation is the distance of the observed value from the true value. But the way distances are measured is determined by geometry and there exists an infinite number of geometries. The errors in statistics are measured as differences between the observed and true values, i.e. by using the Euclidean geometry. The special feature of this (and pseudo-Euclidean) geometry is giving the same weight to all elements of the error. This is not a surprise for a “man in the street”, because the basic education is based on the Euclidean geometry without a warning; this is a very special and not always applicable way of doing things. But even the common sense says that error of observing 1.01 instead of true 1.00 is much more important than “distance” between the bad observations 1.79 and 1.80 of the same true value 1.00. Each element of the error deserves thus its weight dependent on the point of the measuring space. The function determining this way is called *metric*. Euclidean metric has been used in statistics always till the moment, when robustness problems forced them to invent variable weights for error’s elements in the form of M-estimators and other approaches. Different statistical a priori assumptions on data models lead to different ways of measuring errors. But it was Riemann⁷, who in the middle of 19th century declared, that it is not for mathematicians to choose metrics of spaces to model real processes because metrics are given objectively by laws of Nature. Proof of validity of this prophecy was provided after decades by the special relativity theory using the Minkowskian geometry determined by the limitation of light speed and by the theory of gravitation using Riemannian geometries determined by distributions of gravitational masses.

Multiplicity of possible assumptions on statistical data models results in abundance of methods of robust statistics. A question is to be posed if there is a Law of Nature justifying each of these data models sanctifying the applied metrics. However, there is a possibly fundamental problem: is it mathematically consistent to build up the robust statistics using non-linear geometries over the fundament of classical statistics, the roots of which are Euclidean? This problem will be recalled when considering composing of uncertainties.

⁶ Kovanic P., Gnostical Theory of Individual Data, Problems of Control and Information Theory (P.C.I.T.) 13, 4 (1984), 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.

⁷ Bernhard Riemann (1826-1866), German mathematician, founder of the Riemannian geometries.



The first gnostic axiom leads unequivocally to Minkowskian geometry for measuring impacts of uncertainty on data within the quantification process. This is universally valid for all real uncertain data satisfying the accepted model. This confirms the model as an expression of a Law of Nature valid for uncertainty of individual data.

The gnostic binary model of quantification enables the virtual path of the data item's image from its true value to the observed one to be described mathematically and to show its *invariant*: the true value masked by the uncertainty remains constant. Two other features result:

- 1) This path is an extremal of the Minkowskian space: the distance between two points of the quantification path evaluated by the Minkowskian geometry is the *maximum* of lengths of alternative paths.
- 2) When looking for an estimation path from the observed value back to the true data value, the length of which represents the *minimum* of alternative paths, one finds the Euclidean estimation path, the invariant of which is the observed value.

The closed path consisting of quantification path along the Minkowskian circle and estimating path along the Euclidean circle forms the Gnostic Ideal Cycle (IGC).

1.4 The problem of entropy

Risks are measured by probabilities. Uncertainty is a lack of knowledge on states of things or events. It results in necessity to consider probabilities of events instead of sure facts. Development of science in 19th century came to notion of entropy closely related to probability: "Entropy is the logarithm of probability" states the immortal testament engraved on Ludwig Eduard Boltzmann tombstone in Central Cemetery of Vienna. This measure of uncertainty was based on the Maxwell's molecular theory of gas. The chain uncertainty – risk – probability was thus completed by its necessary element. However, the need for entropy originated in the long-time efforts to express the second Law of Thermodynamics quantitatively. The notion of entropy was coined by Clausius⁸ in 1865 for the ratio of heat over absolute temperature. This macroscopic notion was later shown to be consistent with the Boltzmann's statistical notion in its Gibbs' form (1878) accepted by C.E.Shannon as the information entropy. However, there is a serious draw-back of this formula from the point of view of applications, the necessity of a complete probabilistic model of states of the system of variables.

Unlike this, the gnostic formula of entropy of an individual data item has been developed directly from Clausius' formula without references to probability. A Gedanken-experiment⁹ has been performed attaching an amount of energy to an individual data item value to express heat and temperature as functions of the data value. Such an attachment is not strange from the point of practice of analogue computers¹⁰. However, its application respecting the invariance of virtual paths (IGC) passed by the image of uncertain data item during quantification and estimation enabled entropy fields over the data measuring space to be described mathematically. This fruitful result is a unique contribution of gnostic theory unknown from other approaches to uncertainty.

1.5 The problem of information

The generally accepted Shannon's formula evaluating the information was inspired by the Boltzmann's (Gibbs') formula of statistical thermodynamics. It can be formally applied to an individual data item, for which probability of its value is known. However, two obstacles arise:

- 1) How to estimate the data item's probability?
- 2) How to justify application of the formula derived for mass events to an individual event?

Probability of a discrete "random" event could be estimated by observing the frequency of its repetitive occurrence, but not from its single value. To justify applicability of the Shannon's formula for the individual data value, it would require availability of a statistical model delivering the probability. Such a model should work with statistical notions like "population", "randomness" etc. More exactly, such a model must provide the data

⁸ Rudolph Clausius (1822-1888), German scientist, one of founders of the thermodynamics.

⁹ A thought experiment not really performed, but obeying an a priori given system of laws. A famous example is A.Einstein's cosmic lift.

¹⁰ A similar "trick" had been already applied in thermodynamics, too. As pointed out in Baeyer H.Ch.' book "Maxwell's Demon", Random House, New York (1998), both Carnot and Clausius started their thermodynamic investigations by interpreting the amount of gold recovered from a Hindu Kush river's water as energy, gold in water solution as heat and the gold concentration as temperature.



item's probability conditioned on its observed value¹¹. Something like this is not available for an arbitrary data items submitted to analysis.

Both the enigmatic notions of probability and information of an individual data item are derived mathematically in gnostics as consequences of the first axiom by using the already mentioned entropy fields over the spaces of measured uncertain data. (There are two versions of all the formulas, the quantifying and estimating ones). The data item's entropy appeared to be the data weight. Its field's gradient is the irrelevance, the Riemannian measure of the item's error. The formula formally identical with the Shannon's one for a binary system is derived by twice integrating the source of the entropy field (its divergence) along the path of the quantifying/estimating branch of the Gnostic Ideal Cycle. The real version of the result satisfies all Perez's already mentioned requirements to information. The argument of this "quasi-Shannon" formula is a simple function of the irrelevance (gnostic estimating error of the data item), which is interpretable as data item's probability. Both the information and probability (which are real) have their contrasts in complex quantities, which could be accepted as something like disinformation and improbability, because they are rising with uncertainty. Their complex nature should not give rise to astonishment, because physics introduced their wave functions of complex nature long ago when addressing the uncertainty in quantum mechanics.

1.6 The problem of additivity

The composition law of uncertain quantities of classical statistics is additive: a sum of data estimates the arithmetic mean, sum of errors give the bias, sum of error squares estimates the variance, sum of products the covariance. When asked "why?", some statisticians answer "because of using the Euclidean geometry", but this only readdresses the query: why this geometry? An observation and a hypothesis are offered:

- 1) Consider a set of weighted data errors and the tensor of the same number of points of a mass particles rotating around a fixed axis. Attach each data item to a particle by a linear mapping. This mapping will consistently attach masses of particles to weights of errors, momenta of particles to weighted errors and kinetic energies of the particles to weighted squares of errors and sum of non-diagonal tensor's components to covariances. The Conservation Law of classical mechanics composes tensors of movements additively. Then, to make the mapping consistent even for summary tensors, it is sufficient and necessary to compose the errors, their squares and products additively, too.
- 2) The founders of statistics were not only mathematicians, but physicists and astronomers as well. To fit planet's orbit to observed data, it was natural for them to minimize the "energy" of fitting errors under the condition of zero "momentum of errors". The unbiased minimum variance estimate (or OLS, Ordinary Least Squares method) along with the additive composition law of statistics was thus justified by the Momentum-Energy Conservation Law of classical mechanics.

Carl Friedrich Gauss is credited with developing the fundamentals of the basis for least-squares analysis in 1795 at the age of eighteen, but he did not publish the method until 1809. The first publication on this method was that of Adrien-Maria Legendre (1805). It is not sure, that the motivation by classical mechanics mentioned above was anywhere in statistics published; it remains to be a plausible hypothesis.

Unlike this, the composition law of gnostics is non-linear with respect to data and their functions. It is additive with respect to irrelevances and data weights, which are non-linear functions of data. This composition law, accepted as the second axiom of the gnostic theory, is justified by the Momentum-Energy Conservation Law of special relativity theory. This results from the proved linear Lorentz-invariant homomorphism between the quantifying pairs (irrelevance, data weight) and pairs (momentum, energy) of charge-free relativistic particles.

1.7 The problem of optimality

Not an arbitrary estimate is required in data treatment, but the best one. However, what "the best" means, when many optimality criterion exist? The unbiased minimum-variance criterion was already discussed. However, unconditioned minimum variance is also worth of application, when the systematic error is negligible with respect to standard deviation. A useful compromise can be obtained between both¹² estimates. Failures of estimates based on statistical moments can result from unrobustness of estimates.

¹¹ Perez, A.: Mathematical Theory of Information, *Application of Mathematics* (in Czech) 3, 1 (1958) 1-21 and 2 (1958) 81-99

¹² Kovanic P.: Minimum Penalty Estimate (in English), *Kybernetika* (The Cybernetics, Prague) 8 (1972), 5, 367-383



Many other optimality criteria can be found in statistical literature. Some of them minimize the Fisher information. However, to evaluate this information, the likelihood function of the unknown parameter is required, that is not directly obtainable from data and is a matter of the analyst's subjective assumption. Moreover, the Fisher information is not a function of a particular observation, as the considered random variable is to be averaged out.

Following statements result from the gnostic theory:

- 1) Both quantifying and estimating branches of the closed Ideal Gnostic Cycle (IGC) are extremals of the measuring planes.
- 2) When passing these extremals in quantification, the entropy of the data item increases decreasing the information of the data item. This process is out of control of the observing subject, because it is a Law of Nature, which maximizes the "contamination" effect of the uncertainty by using the quantifying path.
- 3) Estimation path is chosen by the observing subject so to minimize this contamination by decreasing the data entropy and increasing its information by using the estimation path.
- 4) The IGC describes an irreversible process: the total balance of quantifying and estimating changes of entropy and information is non-zero for non-zero contributions of uncertainty: the entropy rises and the information falls in dependence on the uncertainty contribution.
- 5) Explicit equation of entropy \rightarrow information conversion as well as of the opposite transformation are derived. They can be interpreted as mathematical model of activity of the Maxwell's Demon, the history of which is described in already cited book of H.Ch. von Bayerer.
- 6) The way of composition of uncertain data is justified by the Momentum-Energy Conservation Law of the relativistic mechanics due to the Lorentz invariant mapping between uncertainty and movement of particles. Uncertainty plays thus the role of the five-th dimension of the movement.

1.8 The theoretical aspects of choosing the data treatment method

To be reliable, a method should be based on a theory. Kuhn (1977, cited in Bird 2007) identified five characteristics that provide the shared basis for a choice of theory: 1. accuracy; 2. consistency; 3. scope; 4. simplicity; 5. fruitfulness. Let us apply these characteristics as criteria for our choice:

The most **accurate** methods are those of mathematics. Methods not based on mathematics can hardly ensure the accuracy comparable with that of mathematics and can be eliminated from the consideration.

Inner **consistency** of a method is warranted by mathematics. But external consistency with Laws of Nature must exist as well. Requirement of consistency with the measurement theory was already discussed. It is given by the fact, that data must be quantified to be treated. But integrity of the Nature and of its sciences leads to requirement of data treatment theory with established natural sciences, especially with recent physics and geometry. This was discussed as well in connection with mathematical gnostics. However, statistics is also an established science deserving the adjective of natural. Alternatives to statistics should also be consistent with statistics in some special cases delimited by the new paradigm¹³. This consistency takes place with gnostics, that shows the convergence of gnostics characteristics of uncertainty to the statistical ones when data uncertainty diminish so that the actually curved data space can be reasonably approximated by a tangential plane.

The **scope** of many methods is limited by assumptions related to the (i.e. statistical) data model. The ideal scope of the data handling method is expressed as the requirement "Let data speak for themselves!". Data are taken "as they are" and all characteristics needed for the treatment are received from data like in gnostic algorithms.

The requirement of **simplicity** must be interpreted as conditioned on the knowledge of the judge. No written text is "simple" for an illiterate. A mathematical proof is generally simple for those knowing the mathematical instruments applied. Mathematical statement is the desired ideal of parsimony.

Fruitfulness of data treatment methods can be measured by applications. To apply a method, it is necessary to go out of scope of a published paper or of a contribution at a scientific meeting. Algorithm must be developed and tested on computers in applications to real data. It is symptomatic of the difficulty of such steps, that only few of the large spectrum of the alternatives to statistics are available in the form of algorithms within the program packages of such computing environments like S-PLUS¹⁴ or R-project. The fruitfulness of the gnostic methodology will be shown in the next chapters in applications and in comparisons with other approaches.

¹³ According to the Czech scientist and writer Bohuslav Blažek (1942-2004), science is developing by explicitation of hidden assumptions. The finiteness of the speed of light was a hidden assumption not only of the Euclidean geometry, but also of the Galilean and Newtonian mechanics. It was removed by A.Einstein.

¹⁴ S-PLUS is a registered trademark of the Insightful Co., Seattle, USA.



It can be thus concluded, that the gnostic data treatment methodology satisfies the theoretical requirements to a new paradigm.

2. RESULTS OF THREE CASE STUDIES

2.1 Problems of testing a data treatment method

Data treatment methods must be tested to show the applicability of the methodology and its theoretical fundament. This can be done by application to simulated and/or to real data. Both choices have their pros and cons.

Artificial generating of data allows the assumed theoretical data model to be warranted. A successful result of the analysis shows then the suitability of the method to treatment of these (may be, only these) data. Worth of such a way is debatable. Mathematics proves its statements mathematically and to try it experimentally can be interpreted as a tautology. The problem of realism of the assumptions remains. However, neither tests by real data cannot be taken as an assurance of method's validity. Mathematical statements are generally valid to all cases satisfying the assumptions. Empirical proofs cannot hold universally.

There also is a danger of subjectivity and "bad will" in empirical testing: when A is testing B's method by application to real data, he can be suspected that his realization of B's method was not enough perfect. To prevent allegation, B's own or B's generally accepted realization or results are to be applied. This is why only the published results and "official" program realizations found in the popular computing environments are used here for comparisons of approaches.

Another problem is connected with unknown values of real uncertain data. Results of different methods differ. How to decide which one is the best? Some ways to solving this problem will be demonstrated in the sequel.

Important problem of testing a data treatment by application to real data is that of the choice of the database. It is desirable to use a typical database enable to demonstrate the features of the tested methods as completely as possible and under hard condition of praxis. To cover the whole spectrum of functions, it is sometimes necessary to use several databases.

2.2 Used databases

2.2.1 Preliminary database

The goal of the deliverable D1.5 of the work-package WP1 was to define the preliminary database for the case study, which should demonstrate the suitability of the methods based on the gnostic theory of uncertain data for risk assessment in health care. As the database suitable for intended case studies, results of a large health survey was selected¹⁵: 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 80 km distance were used as control subjects. The case study I defined in D1.5 was oriented to the robust treatment of data measured below the sensitivity threshold (also called "the left-censored data"). Such a study was completed and its results were

¹⁵ Černá, M. et al., Levels of PCDDs, PCDFs, and PCBs in the blood 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



summarized in the form of a paper¹⁶. Results of the case study II defined in D1.5 were presented in the form of the Power Point Presentation¹⁷ at the Second Annual Meeting of 2-FUN project, Prague, 2-3 February 2009. A review of selected results of this study is presented below.

The preliminary chosen database appeared to be suitable to planned applications. However, there also are limitations in its usage resulting from its narrow concentration on the impact of the people's settlement on their health status. The state of the both surface and underground water was thus reflected indirectly, via the health problems of people consuming products possibly contaminated by pollutants present in water. To extend the scope of the case studies, it was decided to make use of two other databases:

- 1) Results of regular monitoring of the pollutants in Czech and Moravian rivers.
- 2) Results of monitoring the underground water under a dump in Upper Silesia (Poland), which was pursued in the framework of the EU-projects INCORE and MAGIC.

Joint usage of the three databases enabled goals of the case studies to be enriched as shortly described below and as reflected in the deliverable D1.9.

2.2.2 Database from monitoring the rivers of the Czech Republic

The geographic location of the Czech Republic in Central Europe results in its specific role as an “exporter” “delivering” water into several countries: Germany, Austria, Poland and Slovakia. Its contribution to Danube extends its impact into Hungary, Bulgaria and Romania. Along with the river, pollutants of the Czech Republic reach not only the neighboring countries but – finally – three European seas. Such a responsibility gave rise to long-term monitoring of Czech and Moravian rivers. This is traditionally done in cooperation of The Czech Hydro-meteorological Institute with the Institute of Public Health, Ostrava. Available results of this activity during the period 2001-2008 are summarized in the database¹⁸, which contains concentrations of 156 persistent organic pollutants regularly measured at 21 profiles of 17 rivers. Moreover, toxicity effects were also measured by using four methods. This database enabled additional tasks to be formulated for the case studies:

- 1) Interactions between different groups of pollutants.
- 2) Toxic effects of the groups of pollutants.

Results of such analyses were reported at the Prague IPSW09 (International Passive Sampling Workshop) meeting¹⁹ and at the Ad hoc meeting WP3 of the FOKS-project EU and are briefly commented below.

2.2.3 Database from monitoring the groundwater contamination in Poland

The recently running EU-project FOKS (Focus on Key Sources of Environmental Risks, <http://projectfoks.eu>) participated by Germany, Italy, Poland and Czech Republic is a further step of the research activity pursued by the already finished EU-projects INCORE and MAGIC. The database²⁰ summarizes results of a regular monitoring of the water in boreholes surrounding a large dump of wastes originated in mining the coal in Upper Silesia. The worth of this database consists of the following:

- The database is oriented to contamination of the **underground water**.
- The data referred to reflect the **long-term monitoring** (1994-2007) of coherently designed network of boreholes.
- Measurements covered **a broad scale** of 22 chemical and physical parameters of the underground water measured at different depths of the ground.
- Quality of results is **good** enough to provide a rich **experience** useful for similar projects.

¹⁶ Kovanic, P., Ocelka T. and Ciffroy, P., An alternative approach to handle non-detectable values in datasets obtained in environmental and health monitoring, submitted for publication.

¹⁷ Kovanic, P. and Ocelka, T., Loading of POPs in Czech People, Power Point Presentation at the Second Annual Meeting of 2-FUN project, Prague, February 2-3, 2009

¹⁸ DataCHMU_2001-2008.xls, Czech Hydro-meteorological Institute, Prague, 2009

¹⁹ Kovanic, P. and Ocelka, T., Correlations in Pollutants and Toxicities, Power Point Presentation at the International Passive Sampling Workshop, Prague (2009)

²⁰ Trachy 1994 – 2007.xls, project MAGIC



Clever conception of the system of the boreholes allowed many aspects of the contamination process of the groundwater to be investigated as documented in the presentation²¹ and as briefly summarized in the following.

2.3 Analysis of bio-monitoring data

2.3.1 Effects of the left-censored data

The aim of the survey based on the contamination of population was to find out, whether the residents living in the surroundings of the chemical plant are at a greater exposure risk than the controls. Measurement results were subjected to the professional statistical treatment referred to in the cited publication. However, the main tools applied were only point estimates of statistics. They included none probability distribution functions. This approach did not enable the information contained in the high portion of left-censored data to be yielded and the conclusions made to be reliably supported by probabilities of the decision errors. This motivated application of gnostic methods suitable for such purposes.

The left-censored data are those appeared to be below the LOD (Limit of Detection). Instead of the measurement result D representing the value, inequality $D < \text{LOD}$ takes place. There exist several methods of estimation of the true values of such data. The already cited paper compared five of the most frequently used approaches with the gnostic approach based on distribution functions. It showed not only the superiority of the gnostic method, but also its applicability to cases of high percentage of the censored data. Agreement with results obtained by the generalized Kaplan-Meier method popular in applications to survival problems was shown as well. Moreover, unlike the Kaplan-Meier approach, the gnostic method provides estimates of probability not only for some discrete quantiles of the distribution function, but for an arbitrary choice of quantiles including the ultimate point, which delimits the lifetime. This result proved thus the applicability of the gnostic approach to right-censored data case and to survival and reliability problems. The paper also showed the applicability of the gnostic approach to cases of multi-modal data.

2.3.2 Robust testing of hypotheses

Intended tasks for application of the bio-monitoring data included the hypotheses testing. There are significant advantages of the gnostic approach over the statistical one:

- 1) No subjective a priori assumptions on the statistical model of the data are made.
- 2) No a priori assumptions on the kind and parameters of the distribution functions of the zero and alternative hypotheses, everything necessary is determined by the data, i.e. objectively.
- 3) Bounds of data ranges as a by-product of the estimation of the distribution functions are obtained. This enables “almost sure” statements (“with probability 1”) to be derived.
- 4) Not only tolerance intervals, but also relations between distributions are quantified such as the intervals of union and cross-sections of the probability distributions’ domains.

Using the gnostic approach, the main result of the mentioned study of Černá and all. (the effect of citizens’ settling on their health status) could be subjected to quantitative tests. Six pollutants were considered:

- 1) D2378TCDD (denoted P1)
- 2) D12378PeCDD (P2)
- 3) D123478HxCDD (P3)
- 4) D123678HxCDD (P4)
- 5) D123789HxCDD (P5)
- 6) D1234678HpCDD (P6)

Gnostic distribution functions of the global kind were estimated for two kinds of settlements:

- I) The supposed “dirty” place: in the city, where a large chemical plant is located, a former producer of dioxins and other persistent organic substances.

²¹ Kovanic, P., Ground Water Analysis: Experience, Power Point Presentation at the Ad-hoc meeting of the WP3 of the FOKS project, Prague, June 15-17, 2009



- II) The supposed “clean” place in the country in a distance of about 70 km from the plant without any other in a near neighborhood.

The “zero” (“0”) hypothesis is: “There is no significant difference between the settlements.” The “alternative” (“A”) states the opposite. Results of the tests are summarized in the **Tab.1**:

POP	RelQ	LBo	Ubo	LBa	UBa	$\alpha = 0.05$	$\alpha = 0.05$	β	Hyp.”A”
P1	0inA	0.830	2.41	0.205	9.11	$Q < 1.11$	$1.69 < Q$	0.282	Rejected
P2	0inA	0.670	5.15	0.528	5.54	$Q < 1.10$	$2.93 < Q$	0.705	Rejected
P3	0<A	0.011	0.255	0.096	0.453		$0.201 < Q$	0.858	Rejected
P4	0<A	2.0e-7	1.077	7.0e-5	1.346		$0.813 < Q$	0.797	Rejected
P5	Ain0	5.4e-7	0.484	0.00105	0.473	$Q < 0.076$	$0.247 < Q$	0.838	Rejected
P6	Ain0	0.00425	0.948	0.00467	0.369	$Q < 0.031$	$0.345 < Q$	0.975	Rejected

Tab.1: Results of testing hypotheses on health impact of the citizens’ settlements

Following symbols are used in **Tab.1**:

RelQ ... relations between domains of the distributions:

0inA ... domain of the 0-d.f. is inside the domain of the A-d.f.

0<A ... domain of the 0-d.f. is on the left side of the domain of A-d.f.

Ain0 ... domain of the A-d.f. is inside the domain of the 0-d.f.

LBo and **Ubo** ... the lower and upper bounds of domain of the 0-d.f.

LBa and **UBa** ... the lower and upper bounds of domain of the A-d.f.

α ... error of the first kind, (the probability of the “false alarm”)

β ... error of the second kind, (the probability of accepting the hypothesis A, which is false).

Q ... quantile (the value of a data item).

The error of the second kind is high in all six considered cases, which means, that the alternative hypothesis is to be rejected. In other words, there is no significant difference in concentrations of six tested pollutants in blood of the people living in immediate neighborhood of the chemical plant and of those living in a place far away. A pessimistic generalization of this result can be formulated:

“It could be reasonably worried, that there is no “clean” place to live in Central Europe.”

These misgivings about the life in Central Europe were derived from concentrations of six organic pollutants. However, considering the sum of concentrations of all 17 PCDD/Fs congeners and 12 non-ortho and mono-ortho dioxin-like PCBs in 20 citizens of the „clean“ location and testing their contamination against 60 people living close to the chemical plant, one gets a similar result to that of the 6 pollutants: the error of the second kind is 0.355 given the error of the first kind 0.05. Fig.1 illustrates the design of the test.

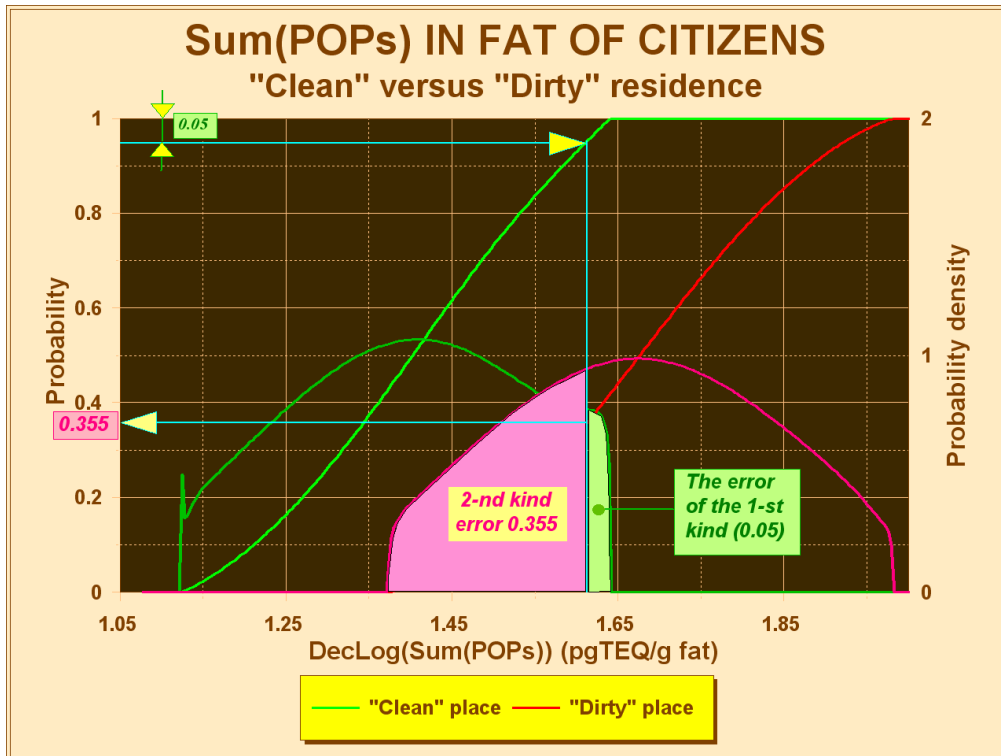


Fig.1: Probabilistic test of the impact of the location of residence on the health of people

Neither a detailed comparison of concentrations does not give rise to an optimism as documented by **Fig.2**. Gnostic distribution functions are inherently robust with respect to outliers. Due to their suitability for data homogenization along with taking in account the left-censored data, they maximize the information obtained. Their mean values obtained by numerical integrating is therefore robust as well. These were evaluated for all measured pollutants and shown in **Fig.2**.

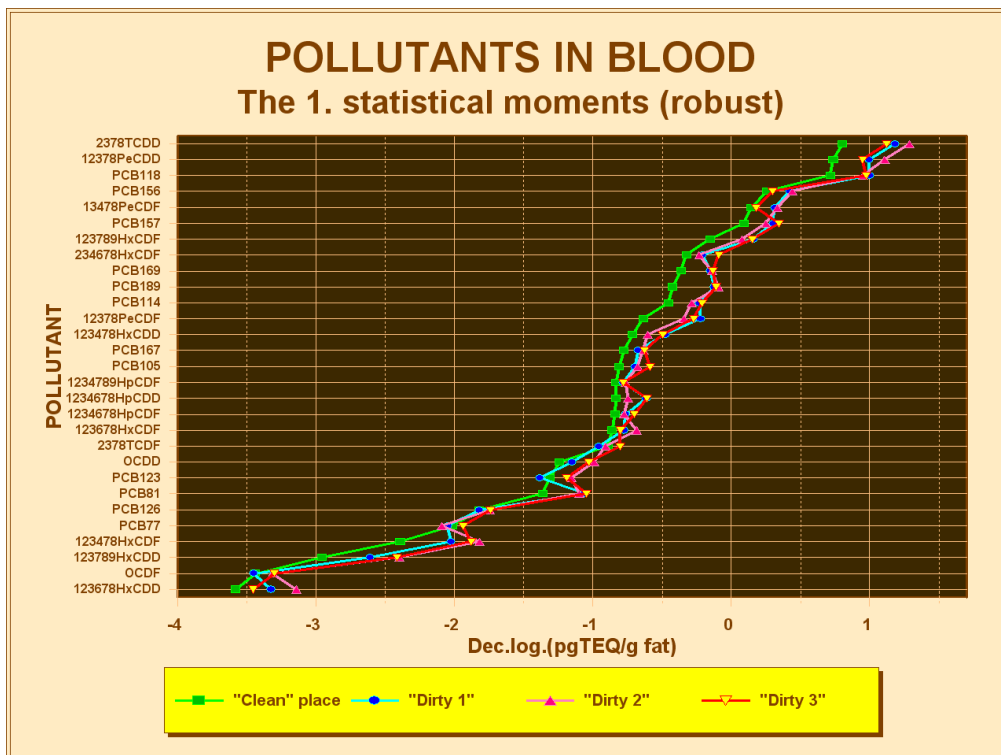


Fig.2: Robust mean values of pollutants in blood of four groups of citizens



Following inference can be based on these figures:

- 1) No substantial differences can be found between the groups of citizens living far from the plant (at the “Clean” place) and those settled at three slightly different (“Dirty”) places close to the factory. (Number of members in all groups is the same, 20).
- 2) The “spectrum” of pollutants’ concentrations is very broad, covering nearly five orders of magnitude.
- 3) The relative form of the “spectra” is essentially similar for each of the four groups.

The last observation allows a daring hypothesis to be formulated: “Czech people are contaminated on a “background”, but substantial level independently on the place of living. Those living close to a chemical plant receive an additional contamination, which “attenuates” (more or less uniformly) the background level.”

One more result can be added to support the scepticism: Take the sum of concentrations of all POPs found (accumulated) in blood of a person and divide it by the age of the person. The result can be called “**the rate of accumulation**”. It will estimate the mean increase in accumulated POPs occurred during a year of the person’s life. Test this rate determined in 20 men and women living in the “clean” place against such rate in 60 citizens settled at the three “dirty” places. The test will say, that the error beta of the second kind (given $\alpha=0.05$) will reach 0.852. In other word, the statement “people living far from the chemical plant are less contaminated” would be false in more than 85% of cases.

For completeness of these results, a look at the relations between distribution functions depicted in **Fig.3** can be useful. These graphs accompany application of the test on the computer. Both probability distribution ($\Pr\{Ho\}$) and its density ($d\Pr\{Ho\}$) of the zero hypothesis (“no difference”) (green) are shown along with those ($\Pr\{Ha\}$ and $d\Pr\{Ha\}$) (red) of the alternative (“significant difference”). The tested variable is the rate of accumulation defined above.

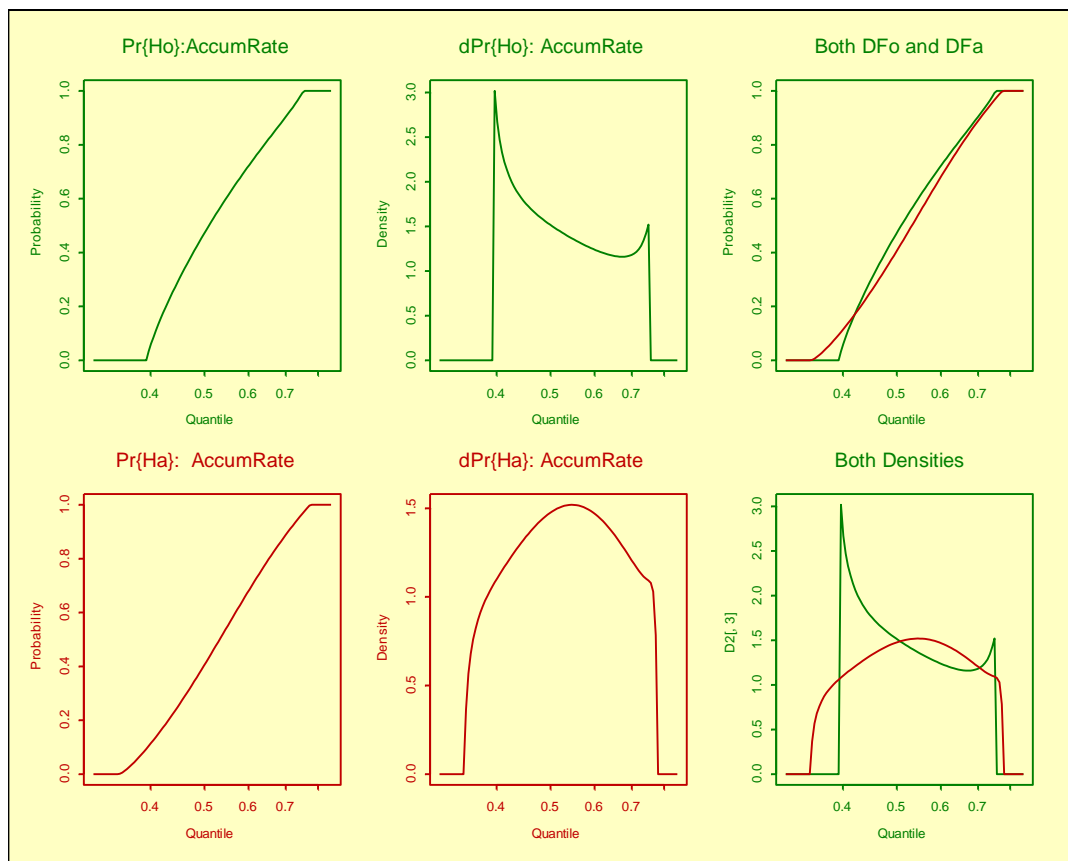


Fig.3: Distributions of rate of accumulation of POPs in a “clean” place (H_0) against the three “dirty” ones (H_a).

Decision making in other problems mentioned in D1.5 can be supported by the distribution functions and quantified by probabilities, but the differences in domains and forms of distribution functions between zero and alternative hypotheses are also too small to be based on standard statistical tests of the kind applied above.

This Fig.3 is instructive among others from the point of view of methodology: it shows, that the distribution functions of these real data cannot be “approximated” by “normal” (Gaussian) distributions.



2.4 Analysis of the surface waters

2.4.1 Correlations in pollutants and toxicities

According to the gnostic theory of uncertain data, a simple (linear) relation exists between covariance and probability of individual items of data samples. These probabilities are estimated by means of gnostic distribution functions, which are inherently robust. Moreover, these enable the information of censored data to be yielded, the homogenization of data sample to be performed and outliers to be objectively identified and eliminated. All this makes the gnostic estimates of covariances and correlations robust and maximizing their information value. To verify these features in application to real data, the monitoring database of rivers were used. Results of this study were described in the already cited presentation “Correlations in Pollutants and Toxicities”. An important step of the analysis is to be mentioned: although the robust estimates were obtained mainly by gnostic methods, results were subjected to statistical tests to quantify reliability of the conclusions. The adjective “significant” applied to the results refers therefore to statistical significance of (mainly) gnostic estimates.

The main results can be formulated in the following way:

- 1) There exist significant correlations between groups of Persistent Organic Pollutants (POPs) as well as between individual pollutants.
- 2) These correlations are mostly positive, but significant negative correlations exist as well.
- 3) It is the group of HCH, which manifests negative correlations with other groups of POPs.
- 4) The gammaHCH is the most significantly correlated substance especially with pollutants of the group PBDE.

2.4.2 Robust models of interactions in pollutants and toxicities

Gnostic robust regression models confirmed the existence of negative correlations:

1. Gnostic results were compared with results of eight methods based on robust statistical theory. It appeared, that gnostic estimates of correlation coefficients were in a good agreement with the geometric mean of eight (differing) statistical methods. More in the next chapter.
2. Four methods of measuring toxicity of each of the six groups of POPs were applied and significance of their results was evaluated by the Test Power (TP , defined as $1-P\{0\}$, where $P\{0\}$ is the probability of zero toxicity):
3. The test power of 0.95 was exceeded by the method Saprobita five times from six cases, while *Daphnia Magna* and *Vibrio Fischeri* were successful in 4 cases from 6 and *Desmodemus Subspic.* only in 3 cases from 6.
4. Groups of POPs manifested significant ($TP > 0.95$) toxic effects in following cases: Σ PAH 4 times from four cases, Σ PCB and Σ HCH 3 times from 4, Σ DDT, Σ PCDD/F and HCB twice from four cases.
5. The best agreement between measuring of toxicities was demonstrated by the correlation coefficient (*Vibrio F.*, *Desm.Subsp.*) equaling 0.524 ($P\{0\}=0.022$).
6. The strongest correlation of pollutants with toxicity was found in the case of (Σ PAH, *Desm.Subsp.*) reaching 0.643 ($P\{0\}=0.010$) and (Σ DDT, *Daphnia Magna*) of 0.501 ($P\{0\}=0.035$).

An interesting question can be posed based on the negative correlation found: Could this effect result from the process of creation the POPs or should it be interpreted as a chemical reaction, an interaction potentially useable to destruction of some POPs? The answer would deserve further investigation

The negative interdependence of some pollutants can be illustrated by graphs of gnostic robust regression models (Fig.4.) The points in Fig.4 show the measured data, the red straight lines are linear models of the dependence of one logarithmic pollutants on the logarithm of the other. The graph in the NW-corner documents the positive correlations between two components of the HCH group. The regression model is of the IWLS-type (Iterated Weighted Least Squares, called in statistics the M-estimator) with the weighting (“influence”) function theoretically proved in literature²².

²² Kovanic, P., A New Theoretical and Algorithmical Basis for Estimation, Identification and Control Automatica IFAC 22 (1986), 6, 657-674



The same method can help in demonstration of the mutually opposite impacts of the pollutants on the toxicity effects measured by the *Vibrio Fischeri* (Fig.5).

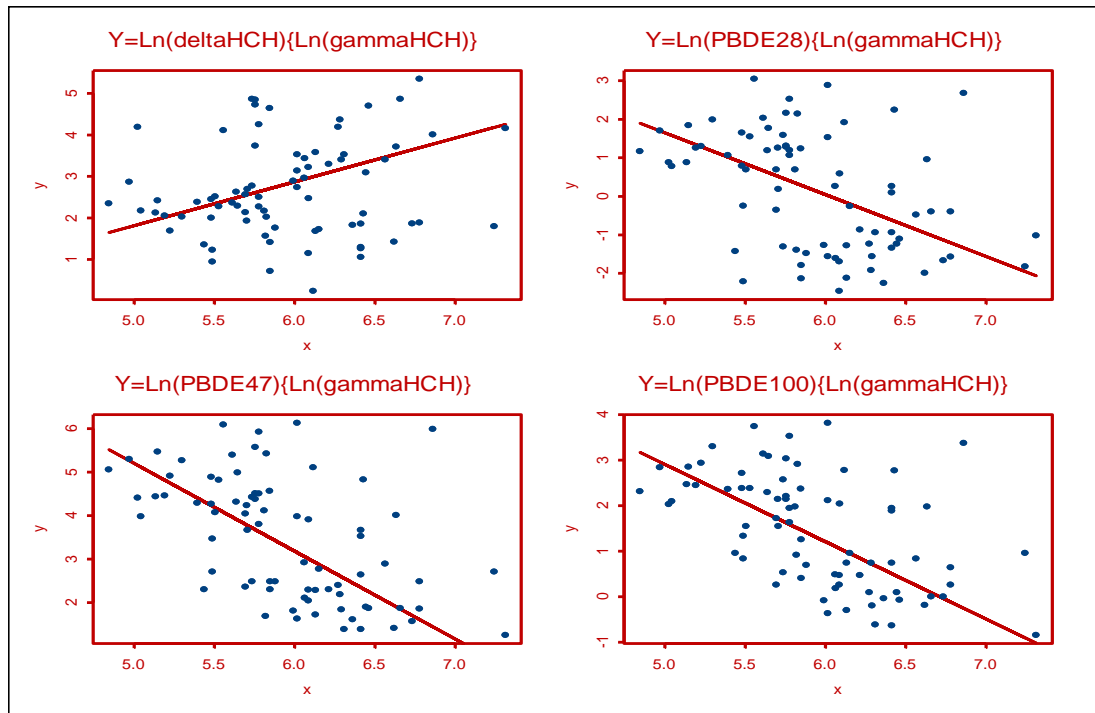


Fig.4: Robust regression models of interdependence of pollutants

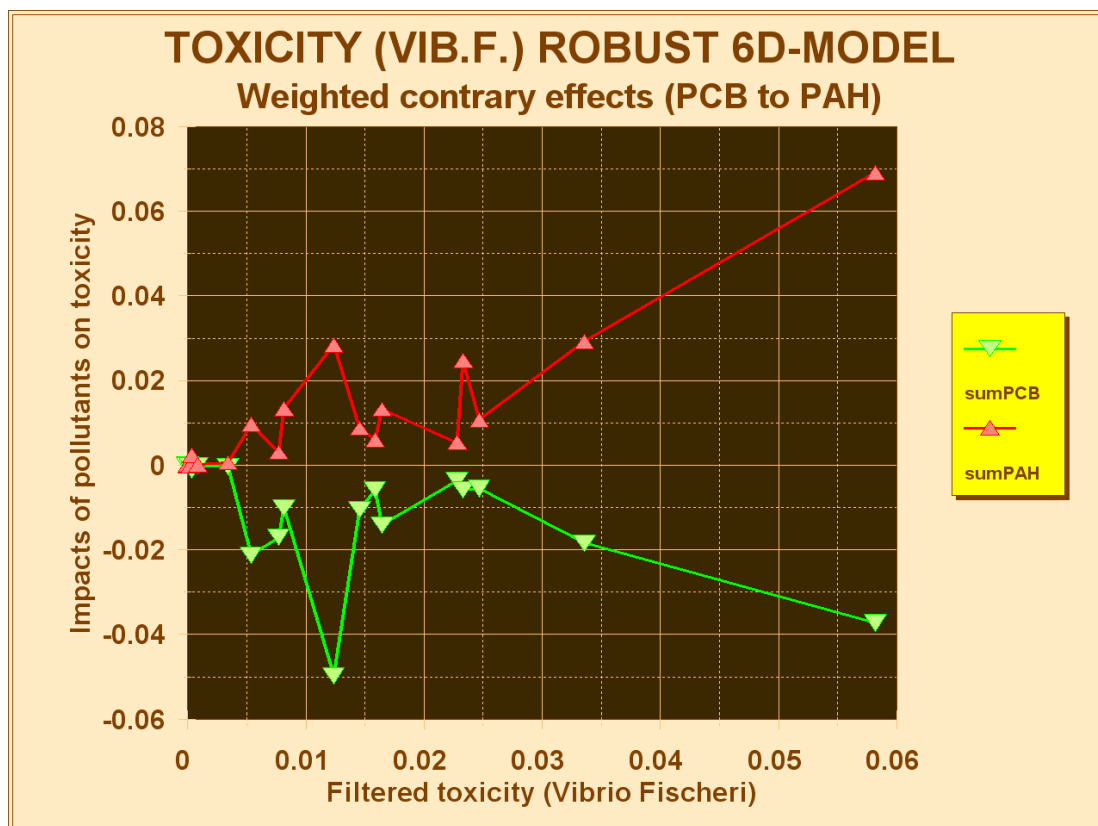


Fig.5: Impacts of two pollutants group on the toxicity *Vibrio Fischeri* in the robust six-dimensional models



The lines in Fig.5 originated in the six-dimensional robust model explaining the toxicity by summary concentrations of six groups of POPs. It should be added, that the correlations of Σ PAH with Σ PCB were positive.

2.5 Analysis of the groundwater contamination

The first worry of an analyst is the quality of the data delivered for the analysis. There are several obstacles from this point of view found in the mentioned database of the groundwater contamination:

- 1) Missing data.
- 2) Uncertainty in data description/definition.
- 3) Uncertainty in data values.
- 4) Data censoring.
- 5) Non-homogeneity of data.
- 6) Outlying data.
- 7) Inconsistency.

The first pair of flaws can be revealed by simple scrutiny and eliminated in cooperation with the data provider. It is frequently necessary to take measures in programs to enable a smooth functioning even in cases of some data missing. Unlike this, the other problems lead to necessary applications of sufficiently sophisticated analytical methods. The groundwater database can be used to demonstrate the range of problems (Fig.6.):

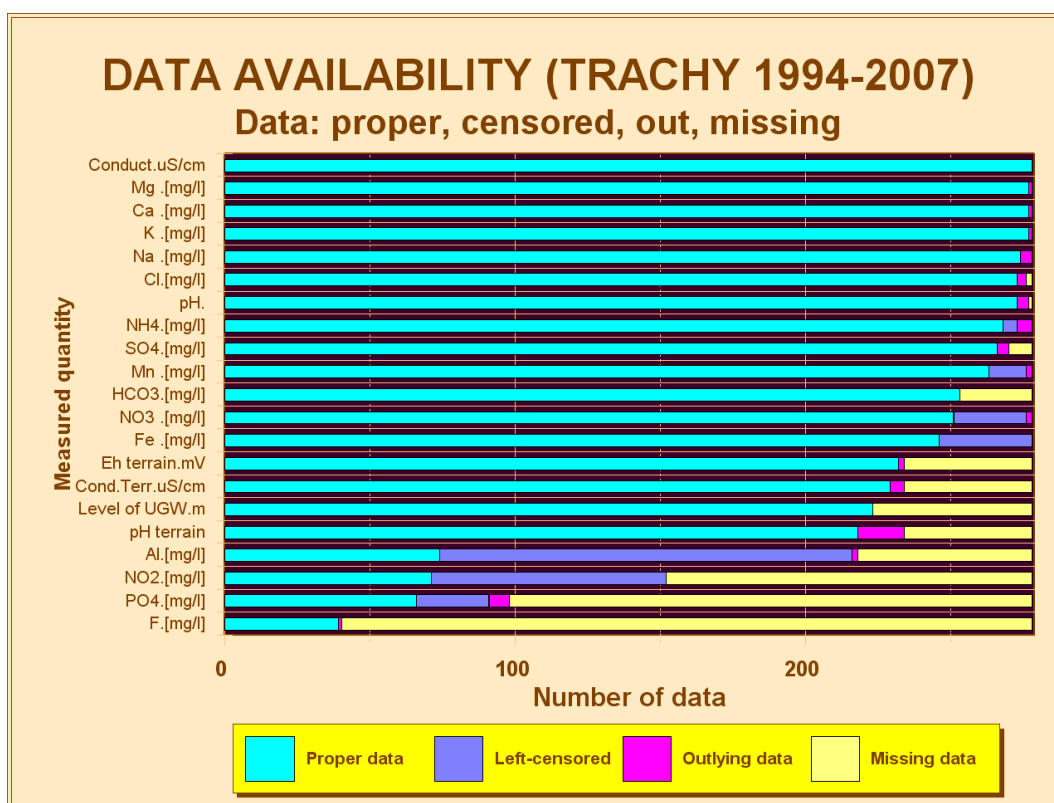


Fig.6: Review of problems devaluating data available for groundwater analysis

The data processing included following procedures:

- 1) Transformation of the data from the form of laboratory protocols into the form suitable to the analysis. Completion of their identifiers (data names, signs of their censoring, proper coding of the missing ones).
- 2) Application of the local distribution functions (ELDF) to reveal the “macro” non-homogeneity.
- 3) Separation of “macro” clusters of data.
- 4) “Micro”-analysis of individual clusters to identify and eliminate the individual outliers, to take out the homogeneous subsamples and estimate their distribution functions along with the bounds of the data range of homogeneous subsamples.



- 5) Robust regression technique to investigate long-term dynamics of the contamination process.
- 6) The same technique to analyze interdependence of measured variables, especially the impact of the groundwater level.

2.5.1 Marginal cluster analysis

Project of the monitoring respected the requirement to evaluate the effect of the dump on the contamination of the groundwater by placing the boreholes not only in points of expected highest contamination (at the south edge of the dump on the north bank of the river) but also in several distant points to measure the background levels. The aim was thus to have two groups of boreholes, one of the “good” ones and the other of “bad”, contaminated boreholes. Results of monitoring enable such an expected classification to be performed objectively. This can be done by the local distribution function *ELDF* (Fig.7.).

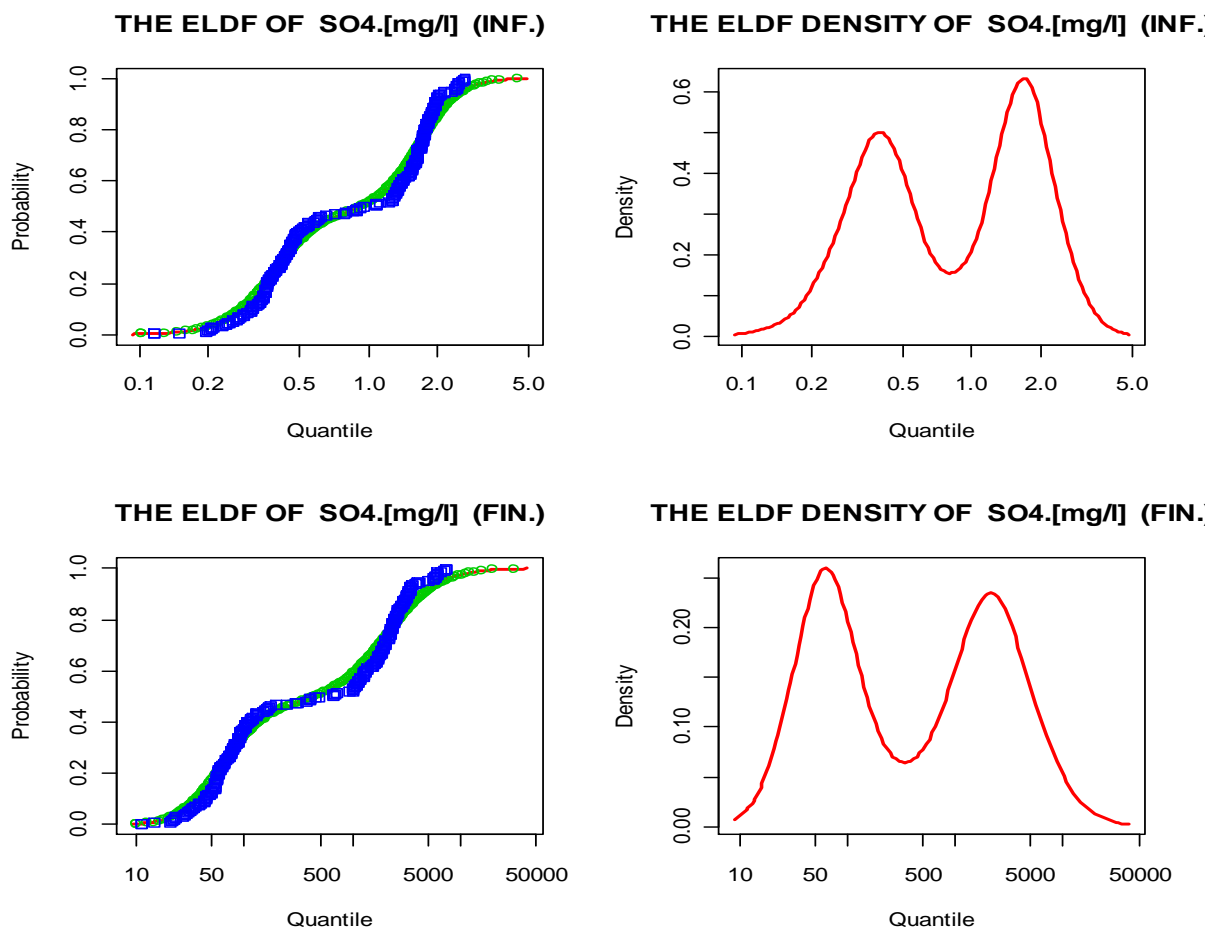


Fig.7.: The local distribution function of the SO4 in all bore-holes

The upper pair of the graphs represents the probability and density distributions over the infinite data support, while the data support of the lower pair is finite, corresponding to the real scale of the data. The advantage of the upper representation lies in its independence of the “boundary” conditions of the real data range (of the lower and upper bounds *LB* and *UB*). The red line is the estimated model of the distributions. Blue squares denote the points of the Empirical Distribution Function directly estimated from data and the green circles are projections of the points of the *EDF* onto the smooth model.

It is evident from the form of the curves that the data sample is non-homogeneous, consisting of (at least) two large data clusters/sub-samples. The gnostic function (*Marganal*) applied provides identification of data belonging to two (or three, if data “say” so) sub-samples.

2.5.2 Homogenization of the sub-samples

The global distribution function *EGDF* is to be applied to sub-samples taken out from the original sample by the function *Marganal*. Bounds *LB* and *UB* of the data support are thus estimated along with the scale parameter *S* and homogeneity test of the cluster with identification of possible outliers. Homogenization by using the function *homogenizeE* is then applied to eliminate the outliers making thus the sub-sample homogeneous. Review of results of application of these operations to all data of the database is in Fig.8. (The redox potential is not enclosed because its value can be negative, not suitable for the scale applied in the Figure).

Three kinds of samples resulted from the marginal analysis:

- 1) Homogeneous samples (HCO₃, Fe, NO₃, Mn, NH₄, Ph of the terrain, level of the underground water, Al, NO₂ and PO₄).
- 2) Nine samples composed of two homogeneous sub-samples.
- 3) One sample containing three homogeneous sub-samples (pH measured in the water).

The light blue lines show the data range of the “main” cluster (having the largest maximum density), green lines belong to the “lower” and red lines to the “upper” clusters. Squares demonstrate the values of bounds *LB* and *UB*. The outliers were demonstrated in Fig.6.

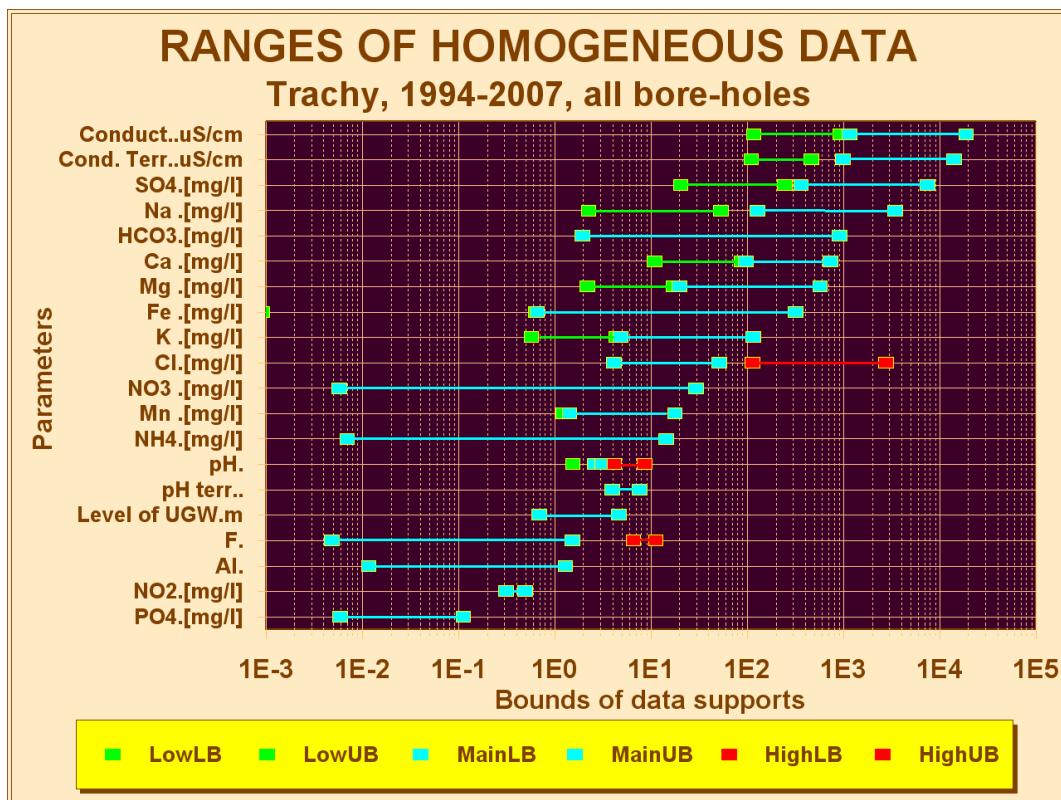


Fig.8: Splitting the original data samples into homogeneous sub-samples

2.5.3 Background and total contamination

To identify the total contamination effect of the dump, it is necessary to compare the contaminated boreholes with the (supposedly) clean ones. As already noted, there were several candidates for the “control” (background) purpose provided. However, reliability of the background measurements performed in different boreholes should be tested. Two boreholes were selected maximally outlying from the dump: P16 in north-east and P8 in south-west from the dump. Their distributions of the strongest contaminant SO₄ were estimated and subjected to the hypotheses test. The lower and upper bounds *LB* and *UB* of both data ranges estimated by the distribution functions were 43.9 and 152.3 for P8 and 46.6 and 111.0 for P16. The hypothesis of “no significant difference between the distributions” was tested by the function *TestHyp*. The error of the second kind (β) was nearly 1.0 for $\alpha = 0.01$. This means, that there is no significant difference between background measurements done in distant points.

2.5.4 The impact of the river on the spread of contamination

The southern edge of the dump lies immediately on the northern bank of the river Bierawka. The underground water flows approximately from north to south. It can be therefore expected that the water in boreholes on the north river bank will be highly contaminated. However, some boreholes placed on the southern river bank enable a question to be asked in what extent the contamination penetrates under the river to the south. To investigate this, three boreholes placed on the north bank (P5A, P3 and P9) were selected along with three southern (P6A, P22A and P8). Results for concentration of SO₄ are reviewed in Fig.9.

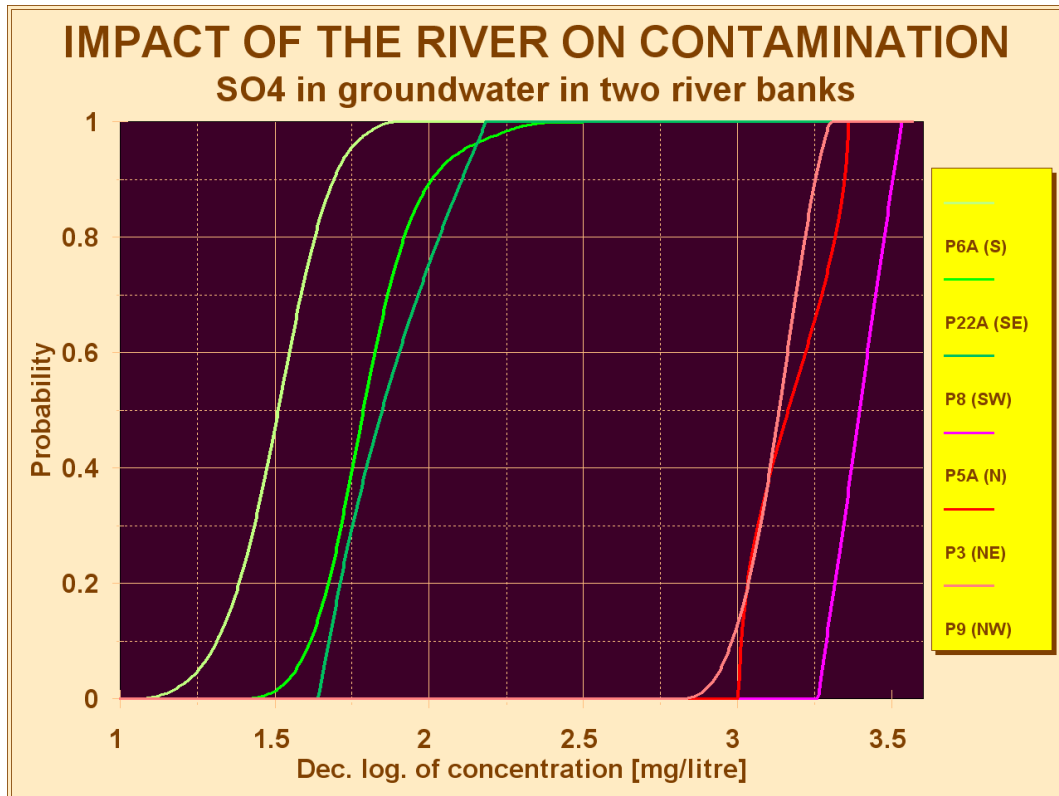


Fig.9: Impact of the river on spread of contamination

The result is unambiguous: The contamination does not reach the southern bank in spite of the difference between concentrations of pollutants on banks reaching orders of magnitude.

2.5.5 Dynamics of the contamination

The considered database originated in measurements repeated roughly yearly over a long time interval 1994-2007. This allows the time changes of contamination to be investigated. Gnostic robust regression procedure was applied to this purpose. It realizes the IWLSQ (Iterated Weighted Least Squares) algorithm known from statistics. The quality of results is dependent on the choice of the “influence” or “weighting” function known in robust statistics in connection with s.c. M-estimators, The form of this functions results from some statistical assumptions on the data model, which should allow to data to satisfy weaker assumptions than those of specific narrow models. Recent software packages contain many procedures realizing such methods of robust statistics. The gnostic weighting function applied in algorithm GWLSX was derived by non-statistical approach applicable without any restriction to data model. The theory of this approach was exposed in already cited paper in Automatica IFAC (1986). The analysis has shown, that the contamination of the groundwater can be substantially non-stationary. An example is demonstrated in Fig.10.:

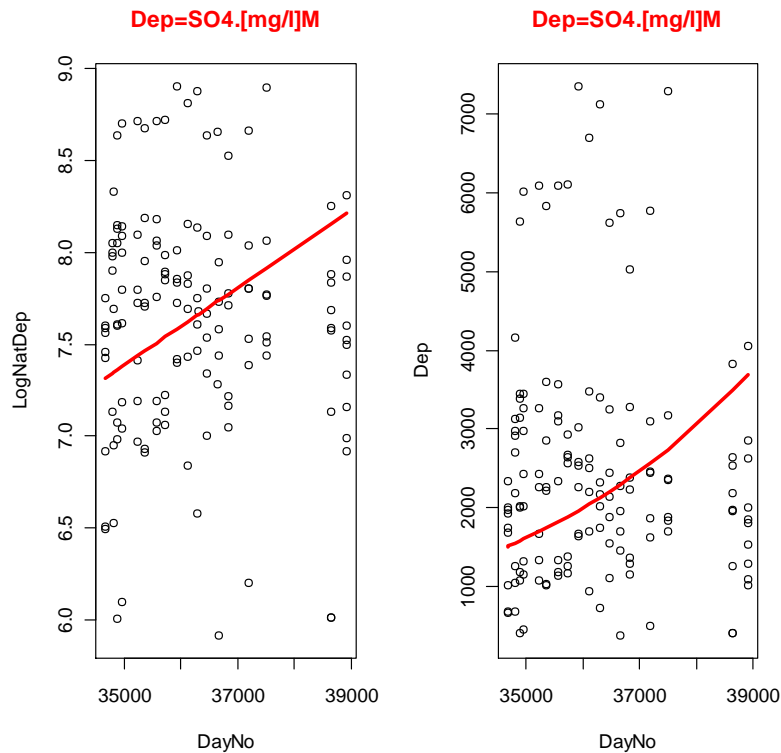


Fig.10: The time-dependence of the contaminant SO4

The difference in graphs is due to the different (logarithmic and linear) scales of the Y-axes. The time on the X-axes is coded as number of days passed after January 1, 1900. The (line) model allowed a quadratic function of time to be applied in fitting the observed data (circles). Significance of the fit was confirmed by statistical tests. Different approaches of robust statistics were used for a comparison, which has shown the superiority of the gnostic approach. Interesting result was obtained by investigating dynamics of different pollutants (Fig.11):

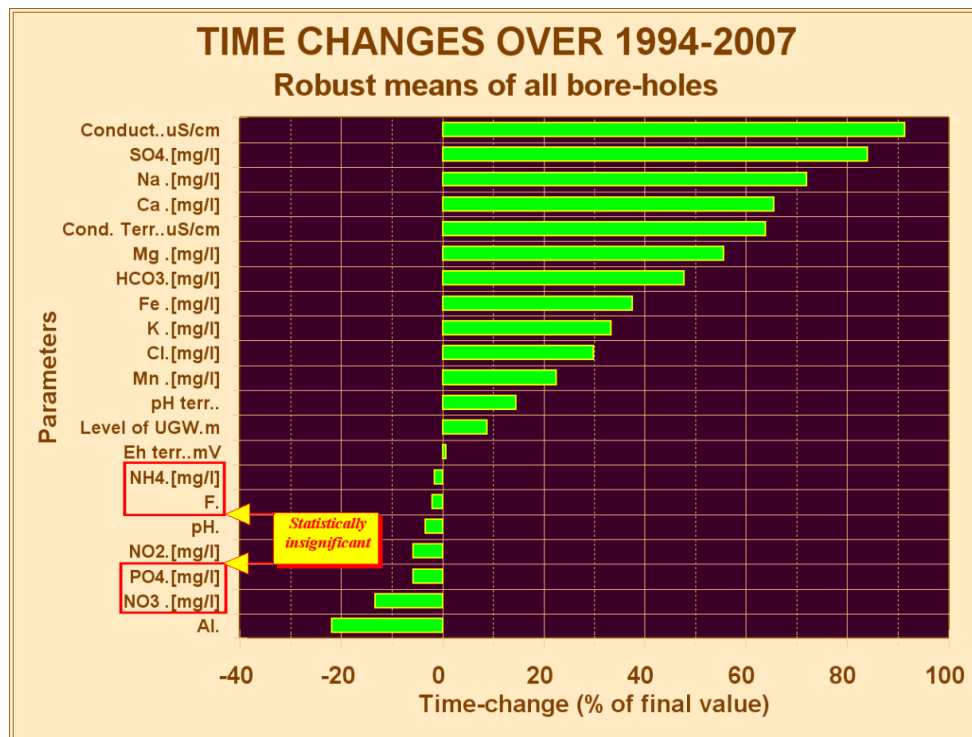


Fig.11: Relative time changes of the observed variables expressed in per cents.



Changes appeared to be insignificant only in four from 21 tested cases denoted by the red frames in the Figure. It results, that different substances and parameters observed change with different velocity, probably dependent on their chemical and physical nature. It also results, that stabilization of contamination is a matter of many years.

2.5.6 Interdependence of observed variables

The same robust regression method was applied to analysis of mutual dependence of the observed variables. Such interdependencies exist, as shown by example of the dependence SO₄(Na) in Fig.12.

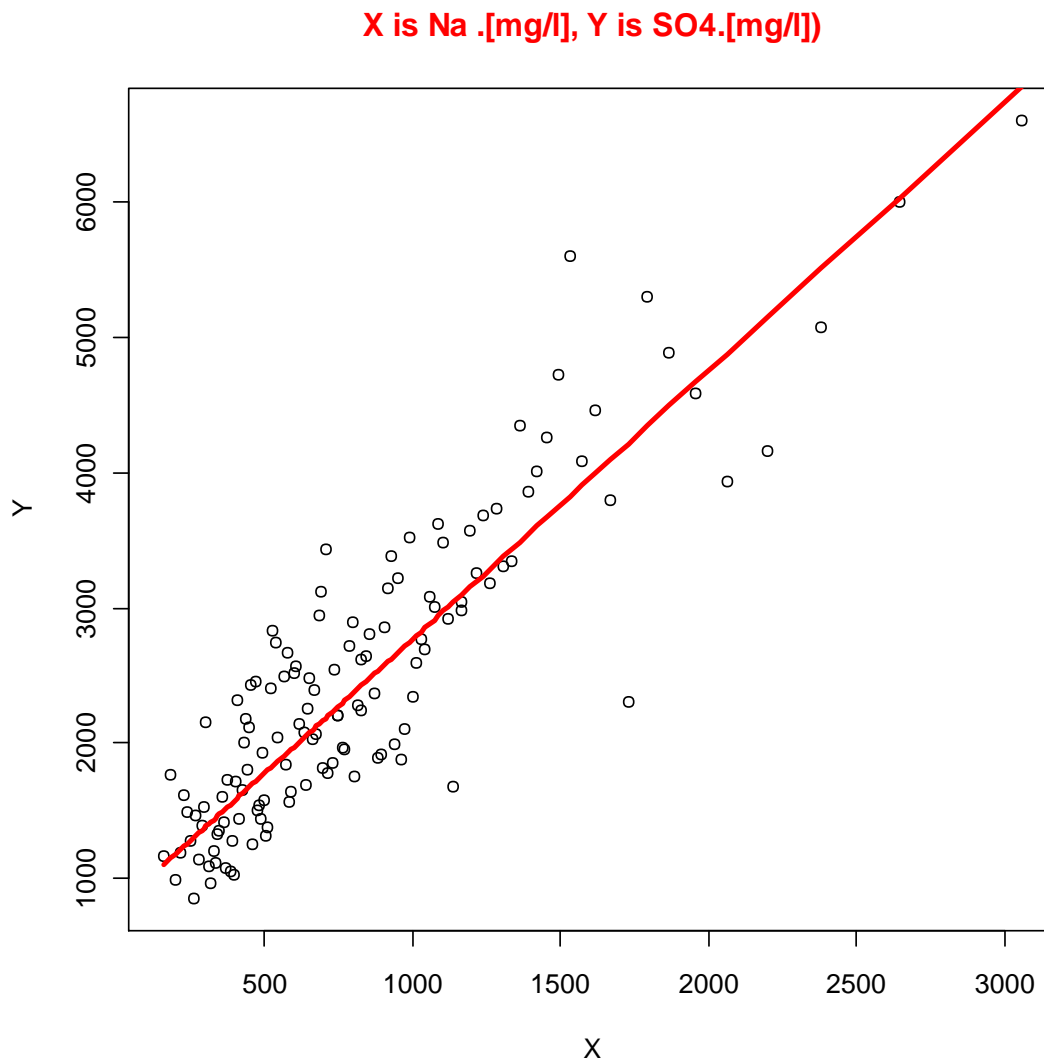


Fig.12: Dependence of the main cluster of concentrations of SO₄ on the concentration of Na in the groundwater

The substantivum “dependence” should be interpreted only in mathematical sense. It does not prove a causal relation of the type “Changes in SO₄ are caused by changes in Na”. Such a linear function can be easily inverted to indicate something like an opposite causal relation. A more neutral expression can be applied: similarity. The sample of SO₄ and that of Na are similar. Only one common factor influencing both pollutants can be excluded, time, because both pollutants were increasing their concentration by a nearly same rate, as seen in Fig.11. The model illustrated by Fig.12 is statistically significant, its R² is 0.974 and P-value (probability of no dependence) is zero.

However, similarities of samples of pollutants could provide a further insight into processes of contamination.



2.5.7 “Scientific water witching”

“Water witching” is a popular way of looking for underground water by using a rod or a forked stick. People’s evaluation of this ancient method is about “forces unknown to science”, “telepathy”, “divine intervention”...

Scientists use in this connection expressions like controversy, scepticism, superstition, myth, charlatanism...

However, the underlying database provides many real data enabling the interdependence between the level of the groundwater and parameters measured on the terrain, at the mouth of the borehole, to be identified. Three parameters of the terrain were measured: conductivity, Eh (redox potential) and pH. Measurements done in “background” boreholes P14, P15 and P16 allow the following relationships to be derived (Fig.13):

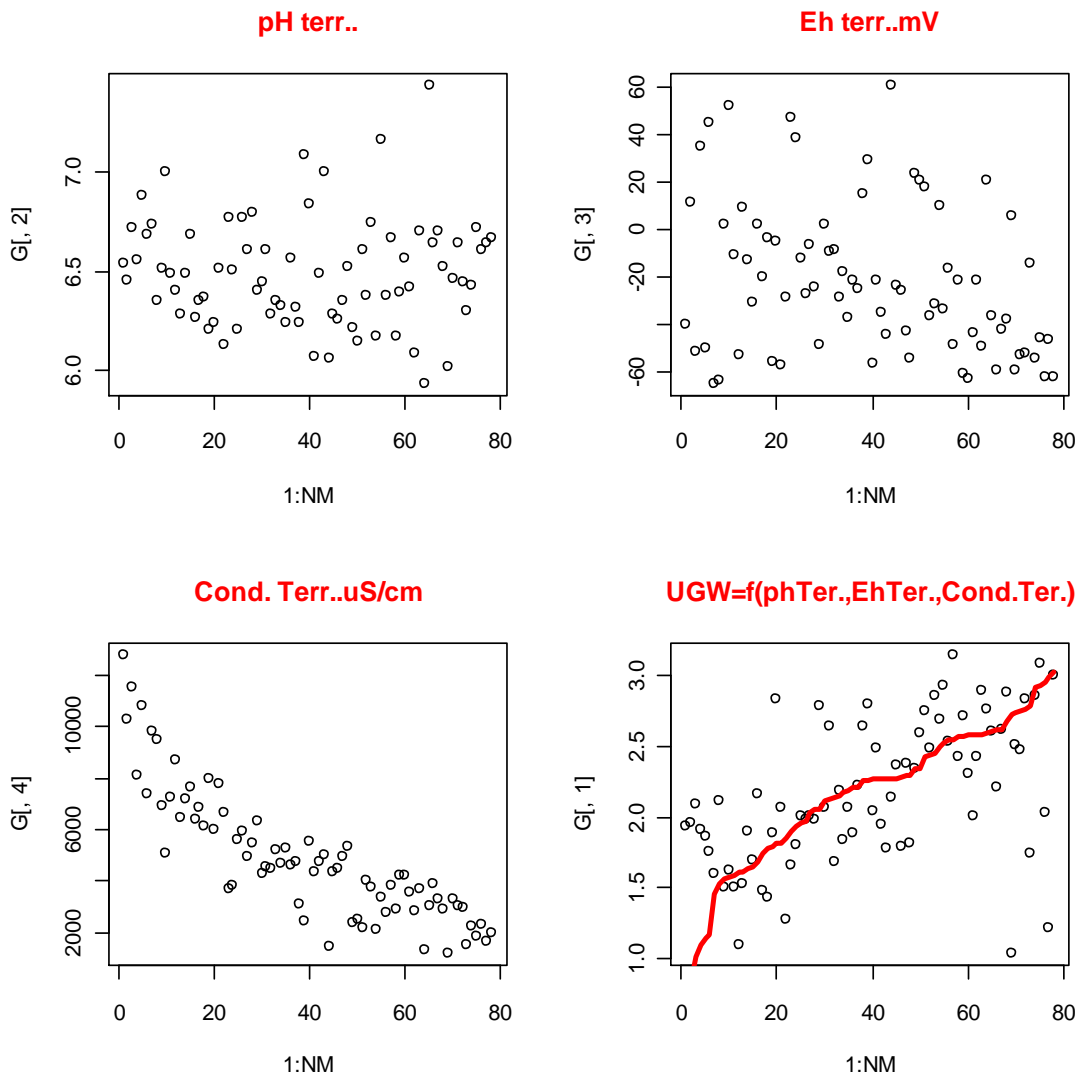


Fig.13: Robust multi-dimensional model of groundwater level as a function of three terrain parameters

The model is statistically significant. Its R² is 0.990. Coefficients of the explaining variables and their P-values are in Tab.2.

	Intercept	pH terr.	Eh terr.[.mV]	Cond. terr.[.uS/cm]
Coefficients	3.175	-0.0291	-7.15e-03	-0.000203
P-values	1.03e-08	0.706	6.7e-16	0.000

Tab.2: Coefficients of the MD-model of the groundwater level and corresponding P-values



Intercept (the constant term), Eh terr. and Conductivity of the terrain are confirmed as significant. Only the pH of the terrain could be left out from the model.

This model proves, that there are no supernatural factors in the ancient “water witching”. Some people are obviously able to sense the physical parameters of the terrain linked with the underground water level. On the other hand, a chance was demonstrated to look for underground water rationally, by means of instruments.

2.5.8 Conclusions of the environmental case studies

Three databases were selected for case studies to demonstrate the applicability and efficiency of data treatment methods based on the mathematical gnostics:

- 1) Bio-monitoring data originated in investigation of POPs in blood of Czech citizens.
- 2) Regular monitoring data on contamination of the rivers in Czech Republic.
- 3) Regular monitoring data on contamination of the underground water by a large dump in Poland.

Following statements can be based on results of the analyses of these data:

- 1) Gnostic method of treating the censored data can be successful even in cases of large portion of censored data in the data samples due to making use of their information in construction of probability distribution function.
- 2) Superiority of the gnostic testing of hypotheses results from its universality of application even to data with unknown distributions. Another advantage lies in estimation of data bounds enabling some test results with probability 1 to be supported.
- 3) Robust testing of differences in distributions of POPs in people living close to a chemical plant and far from it lead to doubts on existence of a place of residence in Central Europe warranting the absence of contamination.
- 4) Robust estimation of correlation coefficients shows both positive and negative significant correlations between pollutants.
- 5) Gnostic estimates of correlations gave results coinciding with the geometric mean of eight methods based on robust statistical theory.
- 6) Gnostic robust regression models confirmed existence of negative correlations between pollutants. They also revealed significant impact of groups Σ PAH and Σ DDT on the toxicity of *Vibrio Fischeri*.
- 7) Gnostic marginal analysis by means of the local distribution function enabled a reliable, robust and sensitive selection of data sub-clusters from the original samples to be performed. Global distribution functions then finished the homogenization by providing robust estimates of bounds of data ranges after eliminating the outliers.
- 8) Robust testing of consistency of measuring background contamination of the groundwater in different boreholes enabled reliable evaluation of total contamination to be completed.
- 9) Analysis of dynamics of contamination of groundwater showed, that the rate of change of different pollutants is very different. Most of pollutants increase their concentration in water significantly with time, showing, that steady state of the dump's impact was not yet reached.
- 10) Robust regression models revealed significant interdependence between contamination parameters.
- 11) A multi-dimensional model can be used to link simple measurements done on the terrain with the level of the underground water.



3. LONG-STANDING EXPERIENCE WITH GNOSTICS

3.1 Applications in economics

The unique suitability of the gnostic theory to evaluate the amount of information obtainable from an individual data item and from small samples of uncertain data enables the information in results of data handling to be maximized. The need for information has been growing at an increasing rate in the modern society and both data and their processing are expensive. This is why the economic aspects of the information must be considered. Following the classical definition of an *economic good* as a good which is scarce relative to the total amount desired and *economic efficiency* as producing such goods at the lowest possible cost, then the idea of a new scientific field called *Economics of Information* is reasonable. Its focus should be oriented on the delivering of the maximum output of information given the cost (of data and of their treatment) so ensure economic efficiency as well. This idea motivated the title of the book²³, which resulted from the cooperation of the Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences with the George Washington University, Washington, D.C. This book not only described the gnostic theory of uncertain data in full detail, but also reviewed in its application part many applications, especially those related to the problems of financial statement analysis, financial management and marketing. This represented a generalization and further development of gnostic applications to economics reflected originally in series of publications and in popular books²⁴.

Following problems were efficiently solved by using the gnostic methods in this field:

- Robust clustering of multi-dimensional models of economic activity of enterprises to decompose several industries into homogeneous groups of firms satisfying the same model. Members of these groups were economically comparable. This enabled the “typical” (recommendable) ratios of economic parameters for financial management to be determined.
- Robust multi-dimensional ordering of firms belonging to the same homogeneous cluster to evaluate the financial position of individual firms objectively.
- Robust multi-dimensional re-appraisal of shares of companies to facilitate rational investment decisions with successfulness reaching more than 70%.
- Robust evaluation of enterprise’s goodwill based on technological parameters of their products.
- Robust gnostic filter applied to data of the computerized inter-banking system of currency exchange tested by real data to show excellent features.
- Robust multi-dimensional monitoring of the financial state of a business to warn on disturbances of the steady state and their causes.
- Robust multi-dimensional recognition of entrepreneurs using the “dirty financing” of their business by keeping a low ratio of turnover of the obligations over the turnover of receivables.

Analyses and their examples were based initially on the data originated in the privatization process of Czechoslovakia. There were continuing by using the industrial data of USA, data from American Exchanges and from international inter-banking system.

3.2 Applications in technology

It is well-known, that the quality assessment systems are traditionally based on simple methods of classical statistics. This conservatism is strong enough to insist on these methods in technical norms, on which certificates of quality depend. This state of things persists in spite of the experience, that these methods cannot effectively solve all problems arising in production. This was why intelligent people working in quality assessment departments and industrial laboratories were open to experiments with the gnostic methods. This enabled

²³ Kovanic P., Humber M.B., *Economics of Information, Mathematical Gnostics for Data Analysis*, (2003), 714 s., to be available on Internet.

²⁴ Kovanicová D., Kovanic P.: *Treasures hidden in accountancy* (in Czech),

- Part I.: How to comprehend financial statements, Polygon, Prague, five editions (1995 – 1998), 256 pp.
- Part II.: Financial Statement Analysis, four editions (1995 - 1999), 303 pp.
- Part III.: Financial control of the growth rate of a firm, two editions (1996 and 1997), 280 pp.



worthful experience from different fields to be obtained along with impulses for improvements and enrichment of the methods:

-
- Quality assessment of ammunition determined by testing shots based on gnostic analysis enabled a high reliability of the tests and their low costs to be reached.
- Consequent analysis of quality of a popular fertilizer led to causes of unstable quality consisting in insufficient control activity of the masters of the storage serving during night shifts.
- Causes of breaking the leading axle of Czechoslovak heavy-duty locomotives exported to Soviet Union were found in unfavorable change of production technology.
- The quality assessment department of a chemical plant producing the caprolactam²⁵ fumbled in vain for causes of not keeping the main quality parameter (transparency) on the required level. Many technological parameters were systematically measured and/or classified, but the decisive one was remaining unknown. Application of gnostic conditioned distribution functions uniquely determined the “wrongdoer” and has shown that control of other parameters was not necessary.
- Tests of fatigue fractures of the springs of the suspension system of heavy trucks TATRA were producing right-censored data. Gnostic methods were successfully applied to choose the best production technology.
- A survey was completed based on gnostic methods connected with the aim of Czech industry to analyze the reliability of nuclear reactors of the pressurized water type working in several countries.
- Surprising contradiction between three classes of physical methods applied in an international geological survey oriented to the cobalt concentration in granit were revealed by gnostic methods.
- Quality of recent electronic scales for weighing continuously running wagons and trucks is determined by transient time of their indication. Gnostic in-line filter resulted in a short time response with no overshoot.

3.3 Applications in monitoring of environment

Gnostic methods were recognized as suitable tools to meet requirements to data analyses in the field of contamination of environment already in 2000. A fruitful cooperation with the Institute of Public Health Ostrava was established. Applicability of gnostic methods to small data samples has been appreciated in the field, where measurements are difficult because of necessity to reveal extremely low amounts of pollutants and where the measuring requires high costs. Moreover, results of these analyses have a vital importance for the society. Many results of regular monitoring of pollutants in waters and in air of Czech Republic were analyzed in this cooperation. This activity also involved applications of gnostic methods within framework of EU-projects MAGIC, 2-FUN and FOKS. Some results of this activity were presented above.

3.4 Other applications

The gnostic theory was originally developed in The Institute of Information Theory and Automation of Czechoslovak Academy of Sciences, Prague. This establishment was focused nearly exclusively on development and application of statistical methods. Development of an alternative to statistics based on scientific methods originated in other sciences (especially in physics) were there neither welcome nor desirable. It represented a deviation from the main line of research. Moreover, tools being applied in gnostics were unknown for statistical professionals who dominated in the Institute. A quite different relation to this “deviation” was that of people of practice knowing the weak spots of statistical methods. Appearance of a method capable to get information from small samples of strongly dispersed data raised a wave of interest of laboratories, research and quality assessment departments of industry. This acceptance provided a feedback to the Academy resulting in “letting the gnostics live”. This interest resulted in development of software suitable for application. On the other hand, requirements of practice served as impulses to further development of the methodology.

Applications to economics, technology and to environmental problems were discussed above. There also were other applications:

- Problems of decision making in health care is connected with fatal aspects. Gnostic methods were used in endocrinology to help in decisions between conservative method and surgical intervention in post-climacterial women.

²⁵ Caprolactam is an intermediate primarily used in the production of nylon 6 fibres and resins.



- A large survey of working stresses performed originally in USA by Prof.Ch.D.Spielberger was continued by Czech psychologists. Application of gnostic methods allowed the stress factors with different professions to be ordered and evaluated.
- Robust models of factors influencing the vitality of living cells cultivated in spinners were tested in a biomedical research project.
- Gnostic methods were and still are broadly applied in the Institute of Theoretical Fundament of Chemical Processes of the Czech Academy of Sciences, Prague, which takes part in international research activity dealing with aerosols.
- Gnostic distribution functions of weights and silver contains of historical coins of Middle Ages enabled interesting facts about the economic life and technology of the historical Czech Kingdom to be revealed.

4. COMPARISON OF ROBUST METHODS

4.1 Comparison of robust location estimators

The initial problem attacked at the dawn of development of robust methods was that of the location parameter of data samples. It was motivated by the high sensitivity of both arithmetical and geometrical means to outlying data. The performances of a choice of eleven statistical estimators of location parameters were compared by S.M.Stigler²⁶ in application to 24 samples of real data. The comparison included sample mean, median, trimmed mean (10%, 15% and 25%), outmean, three types of M-estimators (Huber P15, Andrews AMT and Tukey Biweight), an old estimator of L-type (Edgeworth) and an adaptive estimator (Hogg T1). The data were taken from classical physical measurements: determination of the parallax of the sun (Short 1763), measurement of the mean density of the Earth (Cavendish 1798) and measurement of the speed of light (Newcomb 1882, Michelson 1879 and 1882). Only a part of the data samples (16) were independent. The size of the samples was 17-100. The evaluation of the estimates in Stigler's study was based on the current knowledge of the true values of the physical quantities under consideration: "The closer the realized value of an estimator to the current "true" values of the physical quantities, the better the evaluation of the estimator". Such a point of view resulted in Stigler's conclusions, that "The modern estimators are not worth the time necessary to compute them" and "the smallest nonzero trimming percentage included in the study emerged as the recommended estimator" and "the mean itself did rather well". An extension of the classical testing data using a collection of modern analytical-chemistry data was subjected to a similar comparative study in²⁷. The main point of evaluation of the quality of estimators was that the current "true" value of the classical data is irrelevant to the task of summarizing the measurements because of the possibility of a bias, which may be larger than the variations among data: "What is of importance is the variance of the location estimator used since a lower variance means that the population location parameter is more precisely determined." The comparison of the variance of the estimators applied to extended data samples resulted in the suggestion, that either severely trimmed means or modern robust estimators are required for optimal performance.

Points of view of both publications were retested in²⁸ and compared with the location estimator defined as the mode of the EGDF (Estimating Global Distribution Function) based on the gnostic theory²⁹. The considered data were normalized in the following way: Let E_{ij} be the value of the i -th estimator of the j -th data sample and

²⁶ Stigler S.M. and discussants: Do Robust Estimators Work with Real Data? *Annals of Statistics*, Vol.5, No.6, 1055-1098 (1977)

²⁷ Rocke D.M., Downs G.W., Rocke A.J.: Are Robust Estimators Really Necessary? *Technometrics* 24, No.2, 95-101.

²⁸ Kovanic P., Novovičová J.: On Robust Estimators worth to be Applied to Real Data, Report of the Institute of Information Theory and Automation of the Czechoslovak Academy of Sciences, Prague, No.1463 (1987).

²⁹ Kovanic P.: A New Theoretical and Algorithmical Basis for Estimation, Identification and Control, *Automatica IFAC* 22 (1986), 6, 657-674.



T_j the current “true” value for the j -th sample. Then the quantity $S_j = \frac{1}{m} \sum_i^m |E_{ij} - T_j|$ (where m is the number of estimators) can be used to introduce a new variable $e_{ij} = \frac{\overline{E}}{S_j}$ having interesting features:

1) The set of the e_{ij} represents all the tested measurements.

2) The quantity $RE(i) = \frac{1}{n} \sum_j^n e_{ij}$ (“the index of the relative error” of all n tested estimators) characterizes the set of results obtained by the i -th estimator. The mean of $RE(i)$ for all estimators equals 1.

3) The distribution of the $RE(i)$ is close to normal $N(1, 0.13)$.

To evaluate the estimators’ quality, three criteria were used:

- 1) The mean square error $SE(i) = \sqrt{\left(\frac{1}{n-1} \sum_j^n [(RE(i) - e_{ij})^2] \right)}$ of the i -th estimator applied to all normalized data samples.
- 2) The mean error $RE(i) - 1$ of the i -th estimator.
- 3) The range of errors $r(i) = \max_j(e_{ij}) - \min_j(e_{ij})$ characterizing the distance between the couple of the worst cases not depending on the true value.

Comments are due to these criteria. Normalization of results unified measurements of different physical quantities to interpret them as a large series of m estimates of a unique quantity. The spread of errors is thus a reasonable measure of the estimator’s quality.

Discussion related to the “true” value is necessary in connection with the need to evaluate the mean error. In discussion in Stigler’s paper both professors Stigler and Huber expressed their opinion that the “true” values of the old measurements were the same as the current true values in spite of the discussants Eisenhart, Hoaglin and Pratt pointing out, that the old data were obviously biased because of being obtained by an old measuring technology. But a more reasonable “true” value of the normalized observations can be applied: assume that an “expert board” was called in to express their points of view. The experts would be Professor Mean, Professor Median, Professor Edgeworth and other “authors” of the estimators tested. They would be distinguished by their long serving to the science, experienced, unbiased (at least in the sense of “unprejudiced”) and objective because of using only the data to their judgment. We should believe in their collective wisdom, because we see, that they work “normally”: their judgments have the distribution $N(1, 0.13)$. This is the reason for choosing 1 for the “true” value and the accepted formula for the mean error of the i -th estimator.

The choice of the error range was motivated by the note of Prof. Huber, who expressed his view that the main purpose of a robust procedure is “to prevent the worse”. The error range is an indicator of such an estimator’s feature.

Application of the three criteria to the considered normalized data samples led to results reviewed in Tab.3 in order of the standard error:

	Mean square error	Mean error	Range of errors
Method	$SE(i)$	$RE(i) - 1$	$\max_j(e_{ij}) - \min_j(e_{ij})$
Gnostic	0.038	-0.001	0.139
Hogg T1	0.061	-0.017	0.261
25% Trim	0.070	-0.029	0.261
Edgeworth	0.079	-0.011	0.273
15% Trim	0.104	0.032	0.447
Tukey Biweight	0.131	-0.043	0.631



Andrews AMT	0.147	0.025	0.660
Huber P15	0.210	0.083	0.856
10% Trim	0.211	0.097	0.821
Mean	0.212	0.078	1.055
Median	0.278	-0.124	0.962
Outmean	0.610	-0.086	2.603

Tab.3: Tests of three estimators of location based on the historical physical measurements

It can be thus concluded that the gnostic estimator was the best from all three points of criterial views.

4.2 Comparison of non-parametric distributions

This paragraph is an exception from others, which compare the robust procedures of statistics with the gnostic ones, because the statistical way of estimating the non-parametric probability densities is not robust and there is no other on hand. However, estimation of non-parametric distributions plays an important role in data analysis and cannot be omitted.

Statistical approach to non-parametric distributions of probability based on the Parzen's kernel method³⁰ was described in detail in the deliverable D1.4³¹. Following comments are related to the function *density* available in the R-project environment. The point is that there is much subjectivity in this way of estimating the probability density functions by means of statistical kernels:

- 1) From the seven offered kernel types ("gaussian", "rectangular", "triangular", "epanechnikov", "biweight", "cosine", "optcosine") only the gaussian is differentiable.
- 2) Resulting density is determined numerically in 512 points, but not as a smooth and differentiable function. Formula for estimating the probability density to an arbitrary quantile is not available. It must be substituted by interpolating procedure. Estimating of the mode (densities maxima) and a quantile to an arbitrary density's value is also difficult because of missing differentiability.
- 3) 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, neither which of the provided kernels is the best.
- 4) In case of outlier(s), the resulting density function can unrealistically (and unrobustly) extend the domain of other data.
- 5) There is no recommendation to the optimality of the parameter *width* desirable for a global representation of the data structure. The recommended substitute $bw = "sj"$ does not work always.
- 6) Data bounds (bounds of the data support, of the domain of non-zero density of probability) is not estimated using the data but subjectively set 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*).
- 7) There are tasks for which the (cumulative) probability distribution is needed. To obtain this function, the density must integrated numerically. This does not result in probability as a continuous and differentiable function of an arbitrary quantile.
- 8) There are no tools provided to a direct evaluation of the quality of fitting the true data distribution.
- 9) There are no ways provided for an objective identification of outliers neither for testing the homogeneity of the data sample.
- 10) The function *density* does not assume making use of censored data.

In all of these points, the comparison with the gnostic approach speaks in favor of gnostics. Gnostic distribution functions offer robustness according to the user's wish, either of the external type (with respect to the outliers) or of the internal type (with respect to the internal noises). There is only one, unique kernel, the form (and "width") of which is determined by the data, objectively. Both probability and density are continuous and differentiable

³⁰ Parzen E.: On estimation of a probability density function and mode, Ann. Math. Statistics 35 (1962), 1065-1076.

³¹ Kovanic P., Cifroy P., Review and tests of methods for robust data treatment, deliverable D1.4, project 2-FUN, 29.2.2008

functions. There are two kinds of superposition of gnostic kernels, the global (normalized) and local (additive) ones theoretically justified. The bounds of the data domain are robustly estimated from data, i.e. objectively. Algorithm of the global gnostic distribution function includes procedure for making use of the information of the censored data of three kinds (the left- and right-censored and interval ones). Their estimation results in revealing the outliers and non-homogeneity of the data sample. Optimality of the gnostic distribution is theoretically proved and is related to the maximum of the resulting information.

A comparison of the gnostic kernel with the gaussian may be of interest (Fig.14). The (additive) data on the x-axis were chosen so to expose the whole gnostic kernel (green line). The kernel of the “gaussian” type was then generated for these data using the default parameters of the R-function *density* (the red line). This automatic “width” did not correspond to the gnostic function. The scale parameter was therefore changed to make gaussian’s maximum equal 1 (light blue line). Gnostic kernel gives a larger weight to greater deviation of the quantile from the mean value (0). To demonstrate the difference of both approaches, estimation of density is compared on real data (eruptions times of the geyser faithful, data from the example in R-environment) (Fig.15).

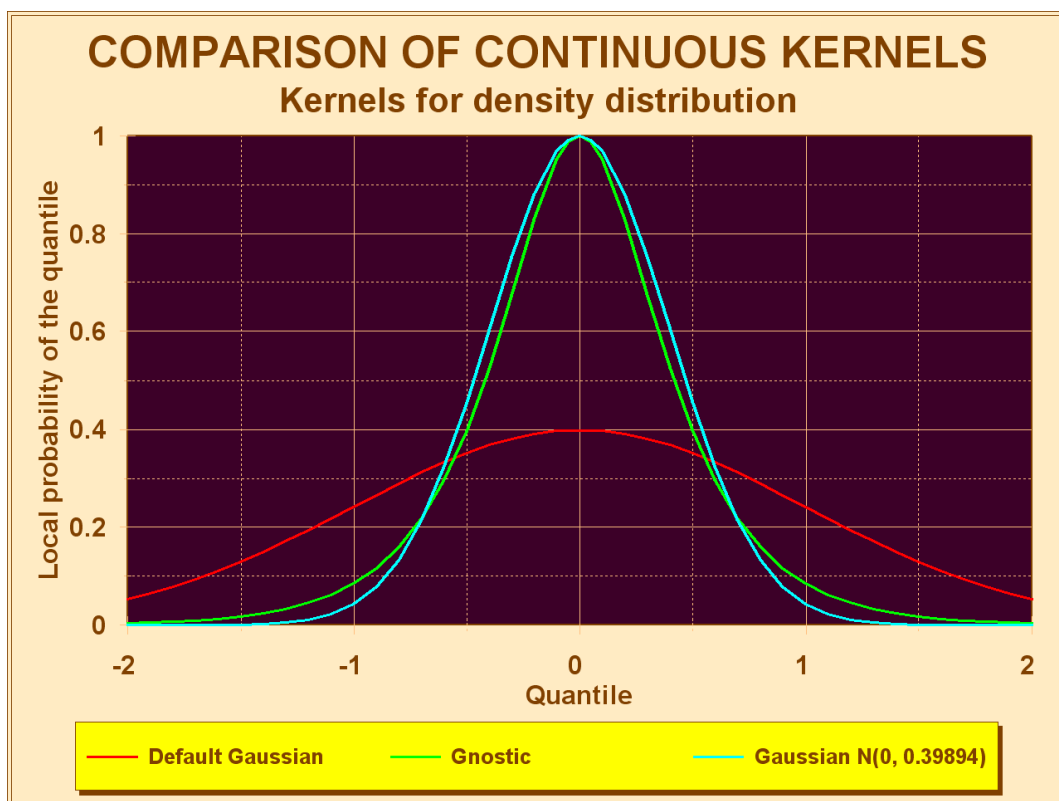


Fig.14: Comparison of the gnostic kernel with the Gaussian

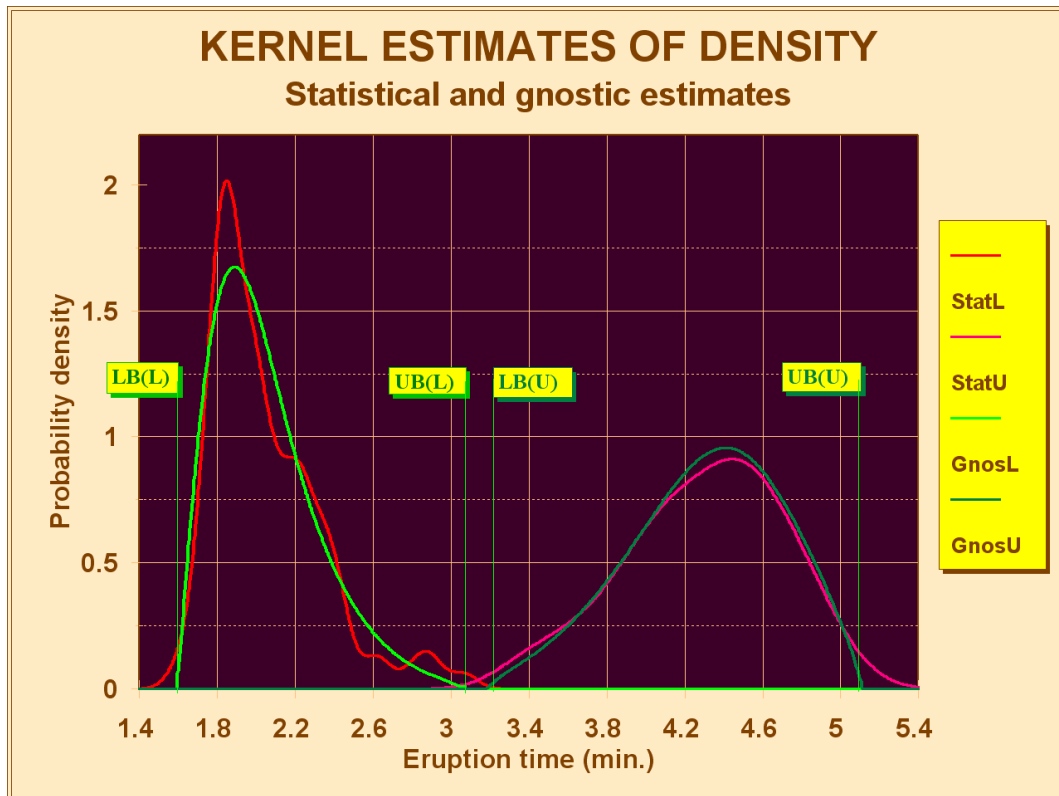


Fig.15: Gnostic and statistical densities of geyser’s eruptions

Statistical density estimated by the function *density* has shown two nearly completely separated densities of two sub-clusters. The minimum between the peaks was reaching density’s value of 0.0296 over the quantile (time) of 3.065 minute. The sample was therefore separated into two sub-clusters, the lower [1.60, 2.90] and upper [3.067, 5.10]. Function *density* with default parameter $bw = "sj"$ (the bandwidth) was then applied to both subsamples (red and magenta lines in Fig.15). Gnostic densities of global distribution functions were then estimated (the light and dark green lines) also providing the bounds of sub-clusters: (LB(L), UB(L)) of the lower and (LB(U), UB(U)) of the upper sub-cluster. The values were (1.595, 3.077) and (3.184, 5.111). This means, that the eruptions are of two kinds with entirely separated domains of probability. Eruption of the length from the interval (3.077, 3.184) have zero probability to occur. This result is supported by data: no eruption lasted more than 2.9 while not reaching 3.076 was observed. Unlike this, both statistical densities estimate the probability over this interval as non-zero. The non-zero probability is also attached to times below 1.595 and over 5.191 by the statistical estimates.

Arbitrariness of using the statistical density estimation can also be demonstrated by Fig.16.

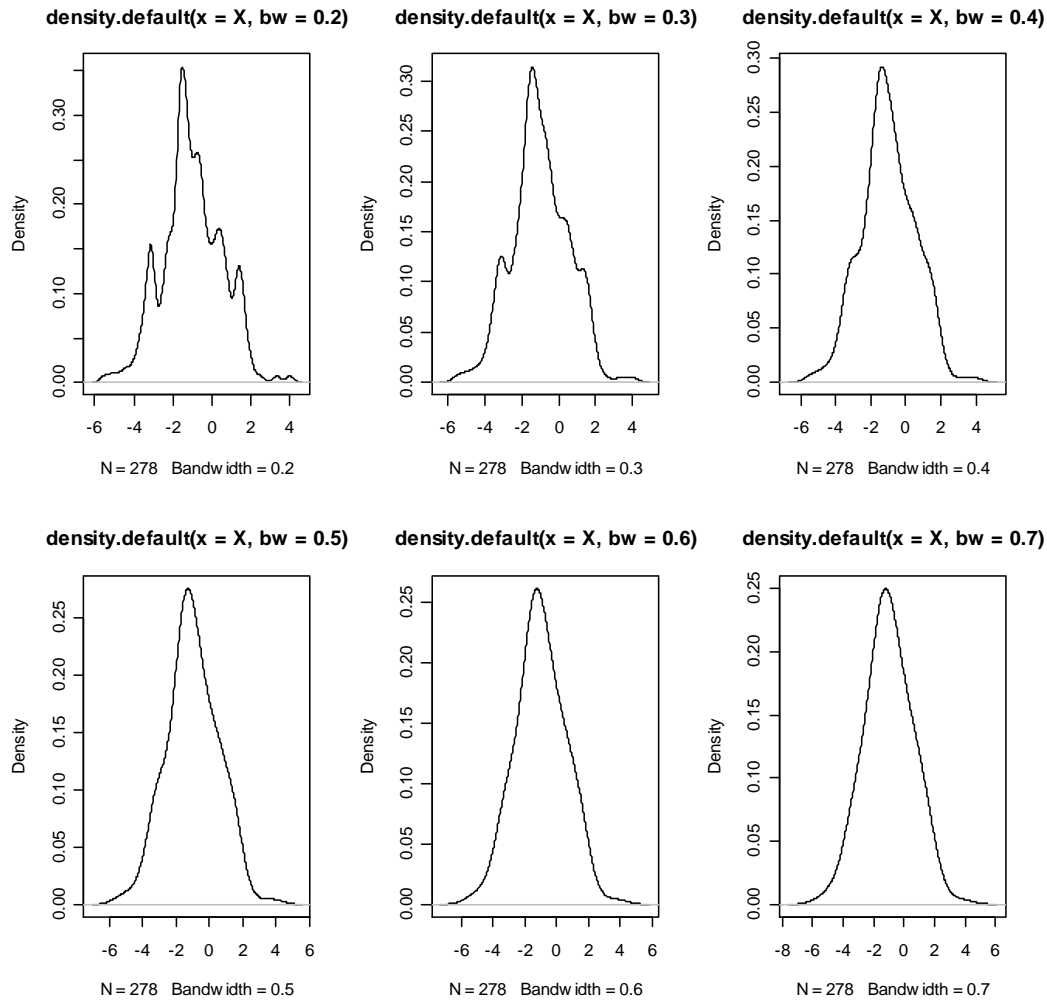


Fig.16: Density of concentration of NO₃ estimated with different parameters *bw*

Data X characterize the contamination of the underground water under a dump by NO₃ [mg/l]. The range of data was extremely broad (0.004, 55.73). Analysis was therefore performed on (natural) logarithms of data. The recommended procedure $bw = "sj"$ failed for these data, it was therefore necessary to try different widths (from 0.2 through 0.7) to obtain density estimates depicted in Fig.16.

Density obtained for $bw=0.5$ is compared with the gnostic estimate in Fig.17. It can be seen from this graph and from Fig.16, that kernel estimation by the statistical way is subjective with uncertainty in adjusting the estimate's parameters. Moreover, it does not provide a reliable information on the finite bounds of the data range.

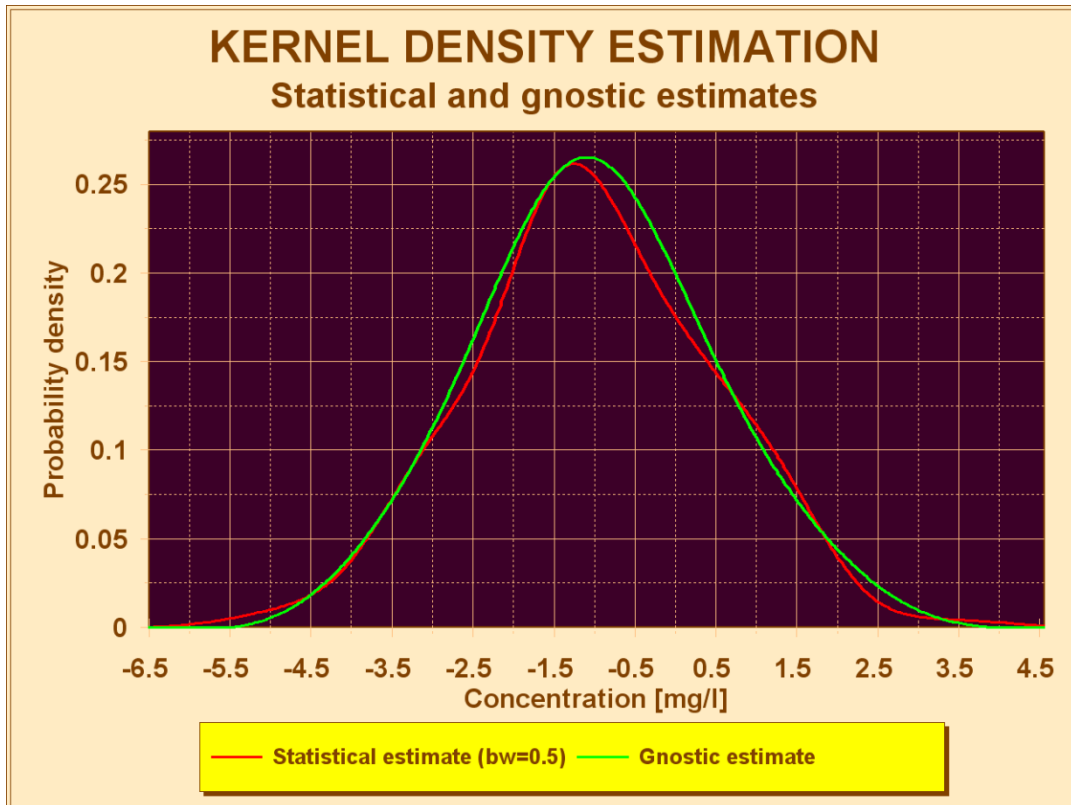


Fig.17: Density estimation of the NO₃ in underground water by statistical and gnostic methods

4.3 Comparison of robust estimates of correlations

Covariances and correlations are popular instruments used to reveal and evaluate similarity of vectors. They play an important role in regression modeling. Both of these statistics used in their original form based on sample estimates of the first and second statistical moments are non-robust with respect to outliers. Their robust estimation attracts therefore attentions of statisticians resulting in a choice of robust algorithms available in literature.

The gnostic estimate of covariance and correlations is based on the simple relation between the irrelevance h ³² of an individual data item and its probability P , which states $P = (1 - h)/2$. This formula is robust with respect to outliers because of the limited range of the irrelevance $(-1, 1)$. The inverse relation $h = 2 * P - 1$ shows that irrelevance – data error – is “centralized” because the zero error corresponds to probability 0.5. Moreover, robustness of the irrelevance is enhanced by estimating the probability P by means of the gnostic distribution function, which filters the data, is robust and uses maximum information of data by treating censored data. It is therefore interesting to compare results of the estimates of correlations obtained by the methods of robust statistics available in the S-PLUS computing system with the gnostic estimates. Application of the officially published realization of the statistical methods contributes to objectivity and credibility of the comparison.

Data used for the test originated in the long-term monitoring of contamination of Czech and Moravian rivers by the persistent organic pollutants (POPs). The question to answer discussed already in 2.4.1 above was if the contamination of different rivers measured by the summary concentration of the POPs was comparable.

Results of applications of eight methods based on the robust statistics and of the gnostic method are summarized in Fig.18.. The largest Czech river Labe (Elbe) and three Moravian rivers (Morava, Odra and Dyje) were tested. It is worth mentioning, that Labe is completely independent of Moravian rivers, its catchment is separated from the Moravia by mountains.

³² The irrelevance h is the error of the data item measured by the Riemannian geometry using the metrics determined by the data.

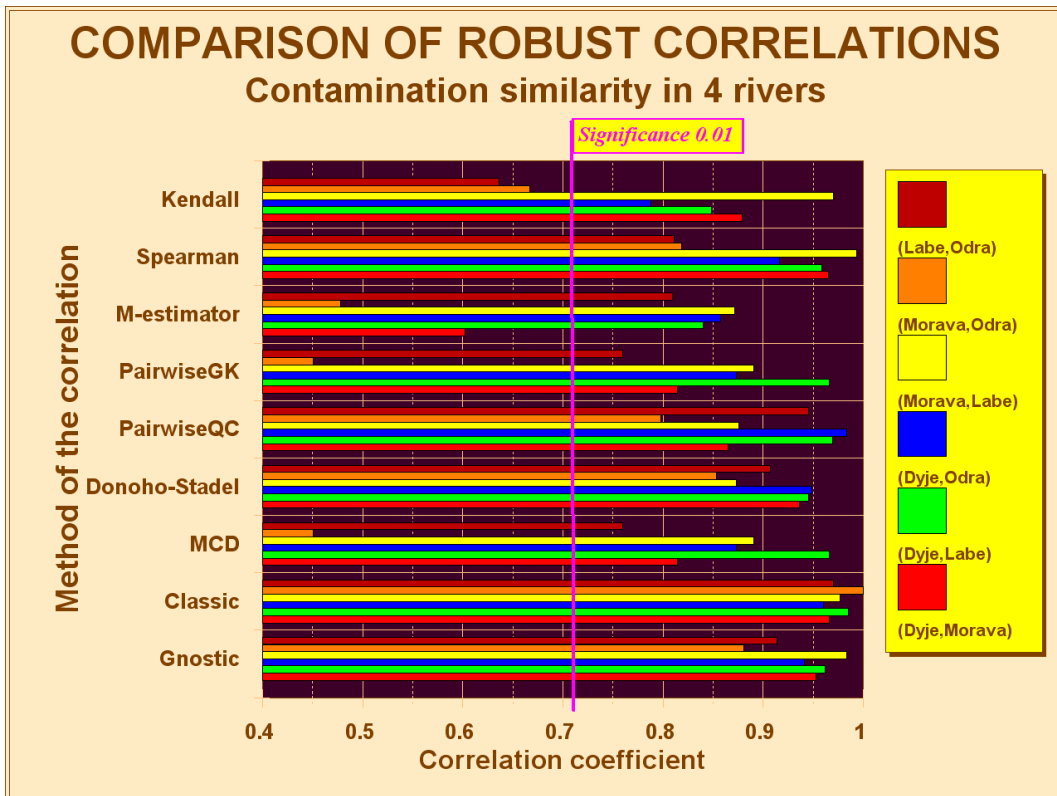


Fig.18: Comparison of statistical and gnostic estimates of correlation coefficients

Several inferences can be drawn from the Fig.18:

- 1) Similarity of contamination of all tested rivers is generally highly significant on the level of $P = 0.01$.
- 2) Different methods give results differing in such a degree, that in 6 from 48 tests done by the statistical methods the significance $P = 0.01$ was not reached.

Two questions remain:

- A) Which of the differing results are true?
- B) Which of the methods is “the best”?

To answer these questions, the “trick” with the “expert board” applied to location estimators can be reused: the blue columns in Fig.19 represent the geometric mean of results of eight statistical methods. When comparing them with the green columns depicting the gnostic results, one can come to the conclusion, that the gnostic method gives results as good as eight statistical methods considered together³³.

³³ This result was presented in the contribution of P.Kovanic and T.Ocelka “Correlations in Pollutants and Toxicities” at the IPSW09 (International Passive Sampling Workshop) in Prague (2009).

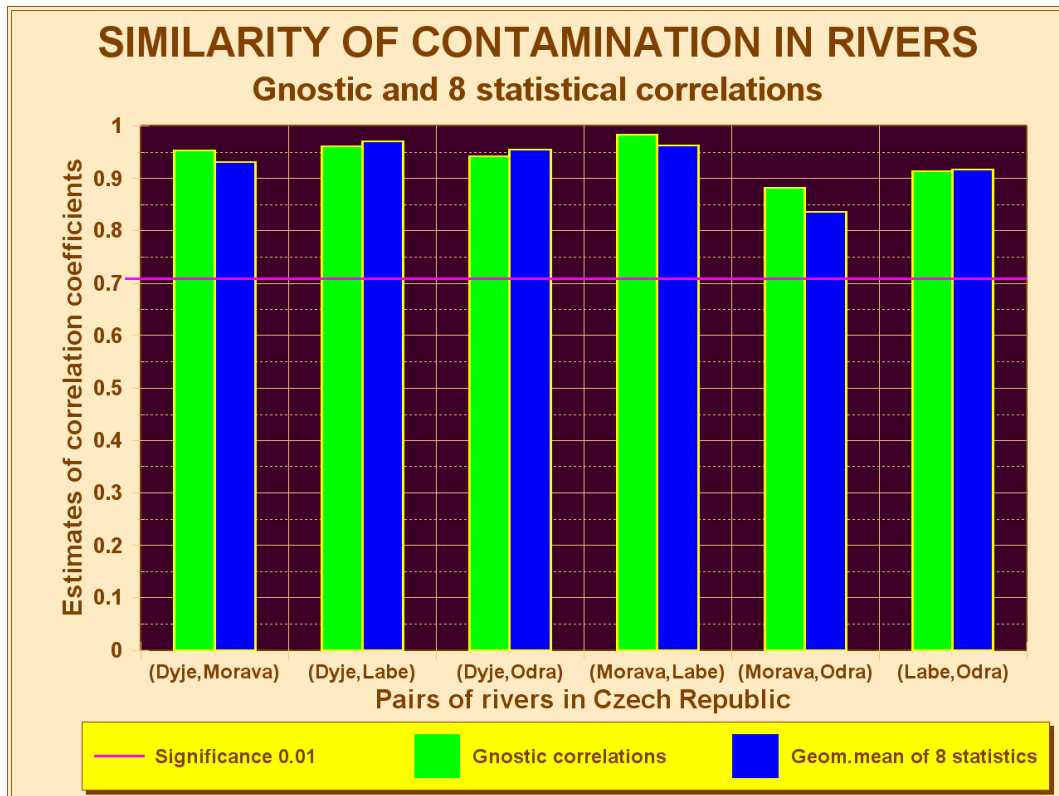


Fig.19: Comparison of gnostic correlations with the geometric mean of eight statistical methods

The magenta line again depicts the significance level of $P = 0.01$.

4.4 Comparison of robust regression models

The classical solution of the linear regression problem, the OLS (Ordinary Least Squares) is unrobust. This motivated the efforts of looking for better approaches. Some of the large number of methods providing robust solutions of the problem are available in the form of algorithms in S-PLUS system, from which those of the class IWLS (Iterated Weighted Least Squares) were chosen for comparison with the gnostic method. There were good reasons for this choice:

- 1) Ten methods of this class are available as functions of the S-PLUS.
- 2) All have some default parameters, which were recommended by the method's author.
- 3) The method IWLS ensures fast and sure convergence.
- 4) This method represents the estimate called M-estimate in robust statistics based on using a theoretically derived weighting function. It was shown in already cited paper in Automatica IFAC (1986), that a weighting ("influence") functions exists based on the gnostic theory, which maximizes information or minimizes entropy of the results of the ISWL method.

The comparison of these methods can be done easily by using an IWLS-procedure with different influence functions. The idea of giving weights to individual equations of the regression model is simple. Weighted mean is a well-known statistics used for treatment data having different spread. In such a case the weight is determined by the ratio of the particular data's variance to the sum of variances of all data. Individual data thus obtain a "collective" and not individual weight, "good" and "bad" data are weighted in the same way. Unlike this, in IWLS method each individual equation of the system receives the weight determined by their "own" error.

Forms of the influence functions are in Fig.20 (those trimming the residues' value or reaching zero for finite residues) and in Fig.21 (continuous functions).

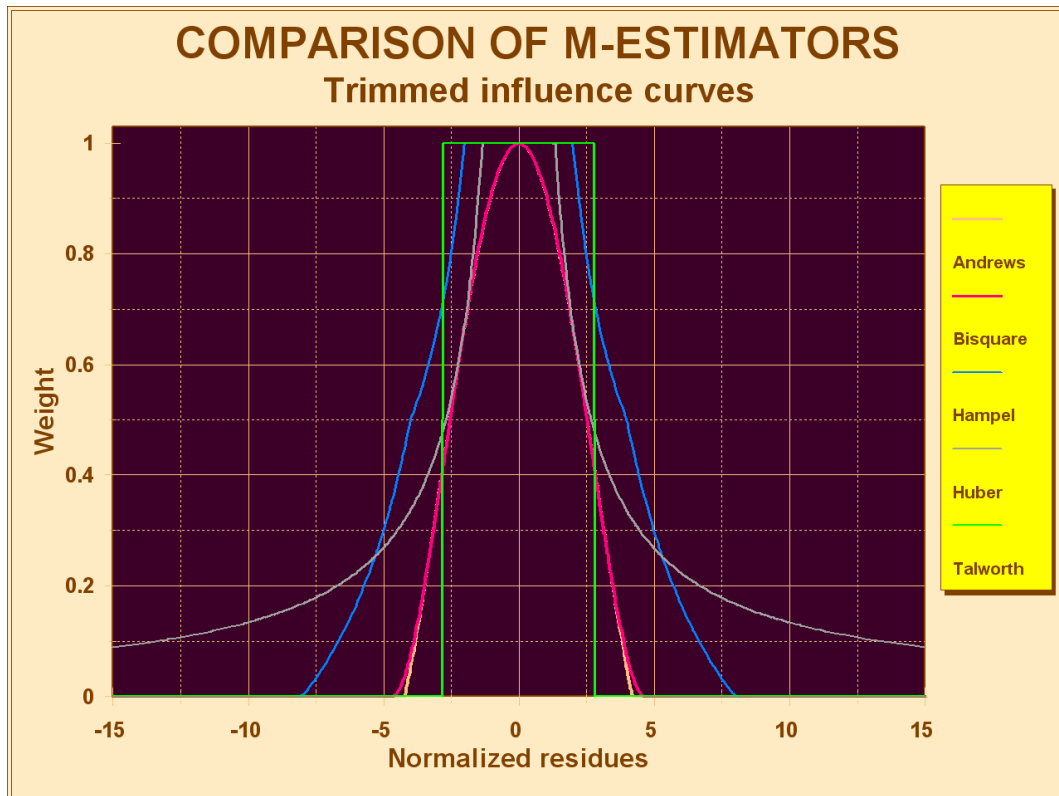


Fig.20: Influence functions non-differentiable over the whole domain

To understand the discontinuing derivatives of the functions, it would be necessary to see the assumptions on which the method was derived. A common sense feeling cannot find reasons why a residue r attaches a non-zero weight while $r + dr$ makes the weight zero. Moreover, trimming is equivalent to ignoring the information in some data. On the other hand, the discontinuity of the derivative can cause problems in operations.

The continuous versions of the influence functions also deserve some comments. The method called “Median” gives – unlike all other methods - extraordinary large weights far exceeding one to small residues. Unlike this, method Fair strongly discriminates residues deviating from zero by very low weights.

The gnostic weighting (magenta line) seems to be almost identical with that of Welsh’s. This coincidence is caused by using the normal distribution for the independent variable on X-axis modeling the values of residues. When applied to a real case of residues, both approaches would work differently, because the Welsch function (like all other statistical influence functions) work with the same scale parameter determined by the formula

$$scale \leftarrow median(abs(resid))/0.6745$$

while the gnostic version adaptively estimates its own scale parameter.

An idea on weights of the estimators under comparison can be obtained in application to real data from the already mentioned monitoring of POPs in rivers of the Czech Republic. Toxicity effects of different groups of POPs were also monitored and investigated by means of four types of methods of measuring toxicities. The variable Tox representing the toxicity effects on the organisms called Fisheri Vibrio was assumed to satisfy equations of the type

$$Tox = C_0 + C_1 \sum PCDD.F + C_2 \sum PCB + C_3 \sum PAH + C_4 \sum HCH + C_5 HCB + C_6 \sum DDT$$

where summing was performed over $j=1, \dots, 19$. The IWLS procedure was run subsequently with the considered M-estimators. The weights reached by the convergence are presented in Fig.21 ordered by gnostic results. It can be seen, that the gnostic estimator was decreasing the weights of seven equations most drastically to give a full weight to twelve “good” ones. Its distinguishing of quality of observations was not supported by the statistical methods. Moreover, the estimator Median attached to seven equations weights from 3260 through $1.14e+10$.

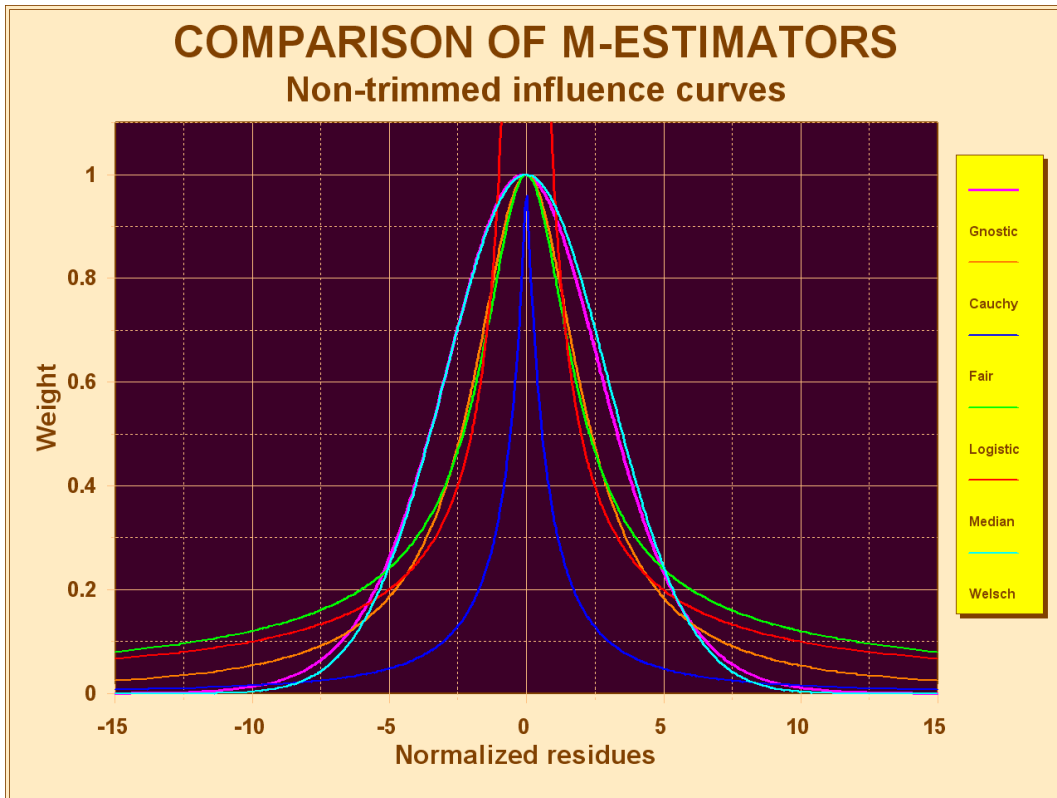


Fig.21: Differentiable influence functions of the M-estimators

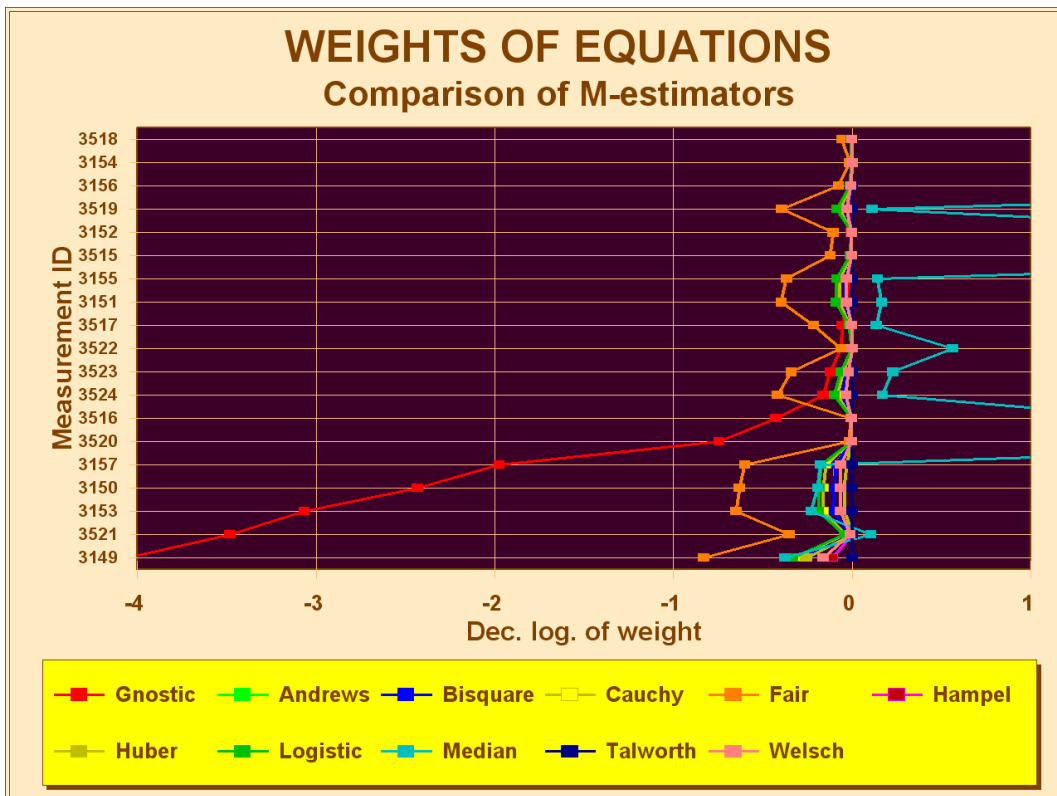


Fig.22: Weights given to equations by influence functions of the M-estimators

There is the statistical function `ls.diag` available in both S-PLUS and R-project computing environment available enabling the statistical evaluation of the weighted linear model to be obtained. The significance of each of the explanatory variables of the model is thus estimated by the *p-value* (probability, that true value of the coefficient of the variable is zero). This value plays thus role of the error of the second kind in testing the significance of the variables. The test power $1 - p\text{-value}$ can be thus estimated for all methods and all explanatory variables. Results are depicted in Fig.23.

Only to one of six considered groups of POPs was the role of the toxic wrongdoer assigned by all methods. Only the gnostic model delivered full test power to all seven parameters of the model. The Median method reached full test power for six parameters, but its results cannot be accepted because of its obviously unrealistic weights given to some observations. This means, that only the gnostic method led to plausible results.

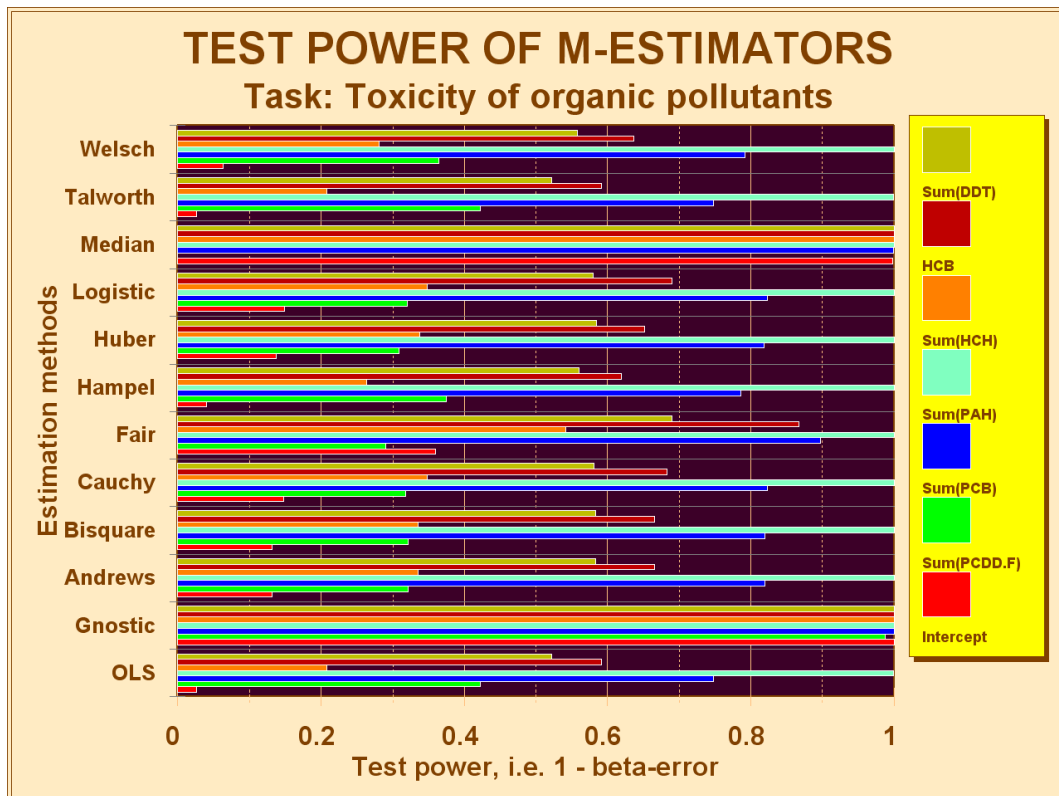


Fig.23: Comparison of the test power of 12 estimators of the M-type

5 CONCLUSIONS

The main idea of including the problem of considering available methods of health risk analysis was to select the method not only satisfying the specific needs of the field, but also meeting the general requirements laid on the paradigm used as the base for solving the problems of vital importance for society. Statistical methods still dominates between approaches of managing the uncertainty in real data. However, problems with this methodology in some applications lead to necessity of finding an alternative especially for cases of data, the amount and quality of which is limited. Methods based on the gnostic theory of uncertainty data are a candidate for playing the role of such a complement and alternative to classical methods. Extended analysis was therefore performed to compare the paradigm of the mathematical gnostics with that of the statistics, to show the applicability of the gnostic methods to different kinds of environmental data, to evaluate their efficiency in case studies, to summarize the long-standing experience with gnostic methods and to compare them with methods based on the robust statistical theory. The justifiability of the paradigm on which is the discussed approach based has been shown. Results of the case studies confirmed its practical applicability. Robustness of gnostic algorithms was demonstrated outperforming that of other tested methods. Methods based on the gnostic theory of uncertain data can thus be recommended for applications to health risk analysis.



6 ACKNOWLEDGEMENTS

Acknowledgements are due to prof. Černá (The Charles University, Prague and The State Institute of Public Health, Prague) for making the database available, to the Czech Hydro-meteorological Institute for their collaboration in monitoring and to the partial support of the EU-projects MAGIC, 2-FUN and FOKS. Thanks belong to the Institute of Public Health, Ostrava, for its non-conservative approach to the considered problems and for its support.